The Ethics of Advanced AI Assistants

Iason Gabriel[*1], **Arianna Manzini**[*1], **Geoff Keeling**[*2], **Lisa Anne Hendricks**[1], **Verena Rieser**[1], **Hasan Iqbal**[1], **Nenad Tomašev**[1], **Ira Ktena**[1], **Zachary Kenton**[1], **Mikel Rodriguez**[1], **Seliem El-Sayed**[1], **Sasha Brown**[1], **Canfer Akbulut**[1], **Andrew Trask**[1], **Edward Hughes**[1], **A. Stevie Bergman**[1], **Renee Shelby**[2], **Nahema Marchal**[1], **Conor Griffin**[1], **Juan Mateos-Garcia**[1], **Laura Weidinger**[1], **Winnie Street**[2], **Benjamin Lange**[2,4], **Alex Ingerman**[2], **Alison Lentz**[2], **Reed Enger**[2], **Andrew Barakat**[2], **Victoria Krakovna**[1], **John Oliver Siy**[2], **Zeb Kurth-Nelson**[1], **Amanda McCroskery**[2], **Vijay Bolina**[1], **Harry Law**[1], **Murray Shanahan**[1], **Lize Alberts**[2,5,6], **Borja Balle**[1], **Sarah de Haas**[2], **Yetunde Ibitoye**[2], **Allan Dafoe**[1], **Beth Goldberg**[3], **Sébastien Krier**[1], **Alexander Reese**[2], **Sims Witherspoon**[1], **Will Hawkins**[1], **Maribeth Rauh**[1], **Don Wallace**[1], **Matija Franklin**[7], **Josh A. Goldstein**[8], **Joel Lehman**[9], **Michael Klenk**[10], **Shannon Vallor**[11], **Courtney Biles**[1], **Meredith Ringel Morris**[1], **Helen King**[1], **Blaise Agüera y Arcas**[2], **William Isaac**[1] **and James Manyika**[2]

[*]Equal contributions, [1]Google DeepMind, [2]Google Research, [3]Jigsaw, [4]Ludwig-Maximilians-Universität München, [5]University of Oxford, [6]Stellenbosch University, [7]University College London, [8]Center for Security and Emerging Technology, [9]Independent, [10]Delft University of Technology, [11]University of Edinburgh

This paper focuses on the opportunities and the ethical and societal risks posed by advanced AI assistants. We define advanced AI assistants as artificial agents with natural language interfaces, whose function is to plan and execute sequences of actions on behalf of a user – across one or more domains – in line with the user's expectations. The paper starts by considering the technology itself, providing an overview of AI assistants, their technical foundations and potential range of applications. It then explores questions around AI value alignment, well-being, safety and malicious uses. Extending the circle of inquiry further, we next consider the relationship between advanced AI assistants and individual users in more detail, exploring topics such as manipulation and persuasion, anthropomorphism, appropriate relationships, trust and privacy. With this analysis in place, we consider the deployment of advanced assistants at a societal scale, focusing on cooperation, equity and access, misinformation, economic impact, the environment and how best to evaluate advanced AI assistants. Finally, we conclude by providing a range of recommendations for researchers, developers, policymakers and public stakeholders.

Our analysis suggests that advanced AI assistants are likely to have a profound impact on our individual and collective lives. To be beneficial and value-aligned, we argue that assistants must be appropriately responsive to the competing claims and needs of users, developers and society. Features such as increased agency, the capacity to interact in natural language and high degrees of personalisation could make AI assistants especially helpful to users. However, these features also make people vulnerable to inappropriate influence by the technology, so robust safeguards are needed. Moreover, when AI assistants are deployed at scale, knock-on effects that arise from interaction between them and questions about their overall impact on wider institutions and social processes rise to the fore. These dynamics likely require technical and policy interventions in order to foster beneficial cooperation and to achieve broad, inclusive and equitable outcomes. Finally, given that the current landscape of AI evaluation focuses primarily on the technical components of AI systems, it is important to invest in the holistic sociotechnical evaluations of AI assistants, including human–AI interaction, multi-agent and societal level research, to support responsible decision-making and deployment in this domain.

Corresponding author(s): Iason Gabriel <iason@google.com>

Contents

PART I: INTRODUCTION

Executive Summary

Iason Gabriel, Arianna Manzini, Geoff Keeling

The development of increasingly advanced artificial intelligence (AI) assistants marks the beginning of a technological paradigm shift. While early assistant technologies such as Amazon's Alexa and Apple's Siri employed narrow AI for tasks such as text-to-speech and intent classification, the emerging class of advanced AI assistants leverage general-purpose foundation models to enable greater generality, autonomy and scope of application. These assistants offer novel services to users, including summarisation, ideation, planning and tool use – capabilities that we anticipate will develop further as the underlying technology continues to improve. Advanced AI assistants thus have the potential for deep integration into our economic, social and personal lives, and could redefine how humans experience and relate to AI.

This paper argues that advanced AI assistants raise a number of profound ethical and societal questions for users, developers and the societies into which this technology is received. These include questions around value alignment, safety and misuse, human–assistant interactions and the broader societal implications of advanced AI assistants – including for equity and access, the economy and the environment. Our aim in this paper is to offer the first systematic treatment of the ethical and societal questions presented by advanced AI assistants, and in doing so to characterise the opportunities and risks of this emerging class of AI technologies.

Six key themes emerge from our analysis:

1. AI assistants have the potential to be a profoundly impactful technology via their deep integration into almost every aspect of our lives. In particular, AI assistants have the potential to serve as creative partners, research assistants, counsellors, companions and even a resource which people turn to when making long-term plans or choosing life goals. As such, AI assistants could radically alter the nature of work, education and creative pursuits as well as how we communicate, coordinate and negotiate with one another, ultimately influencing who we want to be and to become.

2. AI assistants have significant autonomy to plan and execute sequences of actions in line with high-level user instructions. Because of this, they present novel challenges around safety, alignment and misuse. In particular, the more autonomous AI assistants are, the greater the potential for accidents arising from misspecified or misinterpreted instructions and the greater the potential for highly impactful forms of

misuse. To address these potential failure modes, this paper proposes a rich sociotechnical approach to alignment that factors in the needs and responsibilities of users, developers and society.

3. AI assistants may be increasingly human–like and enable significant levels of personalisation. While this is beneficial in some cases, it also opens up a complex set of questions around trust, privacy, anthropomorphism, relationships with AI and the moral limits of personalisation. In particular, it is important that relationships with AI assistants be beneficial, preserve autonomy and not rest upon unwarranted emotional entanglement or material dependence.

4. AI assistants may have significant social impacts, both in terms of the distribution of benefits and burdens within society and by fundamentally altering the ways in which humans cooperate and coordinate with one another. While the failure to coordinate effectively could lead to suboptimal outcomes in the form of collective action problems or other socially problematic situations, cooperative assistants may also be able to identify common ground that was previously out of reach. Given the potential utility of assistants, it is also important that the technology remain broadly accessible and be designed with the needs of different users and non-users in mind.

5. Efforts to properly understand AI assistants and their impact encounter an evaluation gap when studied using existing methods. In the context of AI research, existing approaches to evaluation tend to focus exclusively on model evaluation and are thus potentially less sensitive to more general ways in which AI assistants may underperform when considered as part of a broader sociotechnical system. New methodologies and evaluation suites focusing in particular on human–AI interaction, multi-agent and societal effects are needed to support strong evaluation and foresight in this area.

6. The responsible development and deployment of AI assistants requires further research, policy work and public discussion. On the one hand, AI assistants give rise to a number of novel normative and technical research challenges. For example, questions arise about appropriate privacy norms for assistant–assistant and assistant–human interactions and about how to implement these norms in advanced assistants. On the other hand, developers, policymakers and the public all have a critical role to play in developing and supporting governance initiatives around AI assistants. Building upon wide stakeholder input, these initiatives should aim to develop industry best practice, enable public scrutiny and accountability, and advance policy recommendations and regulatory safeguards that are in the public interest.

The paper has four main sections. Part II introduces advanced AI assistants, in particular defining the technology, explaining its technical foundations and outlining plausible applications. Part III examines value alignment in relation to advanced AI assistants before turning to questions around well-being, safety and malicious uses. Part IV considers a class of ethical questions arising in relation to human–assistant interactions – in particular those concerning manipulation and persuasion, anthropomorphism, relationships, trust and privacy. Part V explores a set of questions at the intersection of AI assistants and society, including questions around cooperation and competition, equity and access, misinformation and economic and environmental impact. It also examines the sociotechnical evaluation of advanced AI assistants. Finally, Part VI concludes with analysis of the underlying themes and with recommendations.

We stand at the beginning of an era of technological and societal transformation marked by the development of advanced AI assistants. Which path the technology develops along is in large part a product of the choices we make now, whether as researchers, developers, policymakers and legislators or as members of the public. We hope that the research presented in this paper will function as a springboard for further coordination and cooperation to shape the kind of AI assistants we want to see in the world.

Chapter 1

Introduction

Iason Gabriel, Arianna Manzini, Geoff Keeling

1.1. The Ethics of Advanced AI Assistants

This paper focuses on the ethics of advanced AI assistants, understood as artificial agents with natural language interfaces, the function of which is to plan and execute sequences of actions on the user's behalf – across one or more domains – and in line with the user's expectations. While AI assistants such as Apple's Siri and Amazon's Alexa have existed for over a decade, our expectation is that more advanced AI assistants, powered by large foundation models, will surpass the capabilities of these earlier systems in a number of ways, including generality, scope of action and overall levels of autonomy. Indeed, the earliest advanced AI assistants, such as Meta AI, Google's Gemini models, Microsoft's Copilot, Inflection's Pi and OpenAI's Assistants API, emerged in the latter half of 2023, and there is good reason to expect rapid increases in generality, scope of action and autonomy as the underlying foundation model technology continues to evolve. If this anticipated trajectory continues to unfurl, advanced AI assistants are likely to raise a number of profound ethical and societal questions for users of the technology, for developers and for society more widely. Taken together, the development of more advanced AI assistants – and their potential for deep integration into our political, economic, social and personal lives – may herald a new phase in our relationship with AI technology; one in which questions about alignment with our individual and collective goals, interests and values come to the fore.

To be clear, a world in which some of us are surrounded by and rely upon advanced and potentially human-like AI assistants, while others do not, may be quite different from the one that we now live in. In certain respects, this world could be a great improvement on the present state of affairs. AI assistants could be an important source of practical help, creative stimulation and even, within appropriate bounds, emotional support or guidance for their users. Practically speaking, efforts are currently underway to develop advanced assistants that are able to function as personal planners, educational tutors, brainstorming partners, scientific research assistants, relationship counsellors and even companions or friends. In other respects, this world could be much worse. It could be a world of heightened dependence on technology, loneliness and disorientation. Although the precise form and capabilities of advanced AI assistants are not yet known, the extent to which tasks may be outsourced to it, the anthropomorphic potential of this technology and the ability to speak to users fluently using human language, all create the possibility of material reliance and unhealthy dependence upon it. The existence of advanced AI assistants may also confer abilities on those who have access to them which are out of reach for those who do not. This could compound the challenge of access and opportunity that we already encounter at the societal level.

Which world we step into is, to a large degree, a product of the choices we make now – and how we choose

to proceed as users of this technology, as developers and as members of the society into which AI assistants may well be received. Yet, given the myriad of challenges and range of interlocking issues involved in creating beneficial AI assistants, we may also wonder how best to proceed. This paper explores a number of deep underlying questions about the ethical and societal implications of advanced AI assistants. By engaging in a practice of robust ethical foresight, our goals are to better anticipate where the tide of technological change may take us and to anchor responsible decision-making as we contribute to, interact with and co-create outcomes in this domain.

The paper starts by considering the *technology* itself and different *types* of advanced AI assistant. It then explores questions around AI *value alignment*, *well-being*, *safety* and *malicious uses*. Extending the circle of inquiry further, we next look at the relationship between advanced AI assistants and individual users in more detail by exploring topics such as *influence*, *anthropomorphism*, *appropriate relationships*, *trust* and *privacy*. With this analysis in place, we consider the deployment of this technology at a societal level by focusing on *cooperation*, *misinformation*, *equity and access*, *economic impact* and *environment*, and we look at how best to *evaluate* advanced AI assistants. Finally, we conclude by providing some further reflections on what we have found.

Ultimately, AI assistants that could have such a transformative impact on our lives must be appropriately responsive to the competing claims and needs of users, developers and society. Moreover, their behaviour should conform to principles that are appropriate for the domain they operate in. These principles are best understood as the outcome of fair deliberation at the societal level, and they include laws, norms and ethical standards.

1.2. Key Questions

The span of questions raised by advanced AI assistants is wide-ranging and potentially daunting. In this section we provide an overview of some of the key questions that arise in this context. Each question receives detailed treatment in a later chapter dedicated to the specific topic. The intention of this section is only to provide some sense of the wider ethical landscape – and of the underlying motivation behind this paper.

This overview may also be helpful to readers because of the interlocking nature of the challenges and opportunities that advanced AI assistants give rise to. Awareness of one set of issues frequently feeds into and supports deeper understanding of another. In total, we present 16 clusters of questions about advanced AI assistants relating to the deeper analysis and themes that surface in this paper. The full structure of the paper and chapter contents are covered in the penultimate section of this chapter.

Key questions for the ethical and societal analysis of advanced AI assistants include:

1. What is an advanced AI assistant? How does an AI assistant differ from other kinds of AI technology?
2. What capabilities would an advanced AI assistant have? How capable could these assistants be?
3. What is a good AI assistant? Are there certain values that we want advanced AI assistants to evidence across all contexts?
4. Are there limits on what AI assistants should be allowed to do? If so, how are these limits determined?
5. What should an AI assistant be aligned with? With user instructions, preferences, interests, values, well-being or something else?
6. What issues need to be addressed for AI assistants to be safe? What does safety mean for this class of

technologies?

7. What new forms of persuasion might advanced AI assistants be capable of? How can we ensure that users remain appropriately in control of the technology?
8. How can people – especially vulnerable users – be protected from AI manipulation and unwanted disclosure of personal information?
9. Is anthropomorphism for AI assistants morally problematic? If so, might it still be permissible under certain conditions?
10. What are the hallmarks of an appropriate relationship between human users and advanced AI assistants? When is a relationship inappropriate and why?
11. How should AI assistants interact with one another? In what ways might interaction failures lead to social harm? Conversely, what kinds of benefit might successful cooperation unlock?
12. How might the introduction of powerful AI assistants affect the relationship between users and non-users? What forms of inequality do we need to countenance and address ahead of time?
13. How are advanced AI assistants likely to affect the information ecosystem and public fora? Will they compound or ameliorate the problem of misinformation and disinformation?
14. How are the economic benefits and burdens created by AI assistants likely to be distributed across society? What can be done to ensure that benefits are distributed widely?
15. What is the environmental impact of AI assistants likely to be? What can be done to ensure that their future adoption is compatible with global climate goals?
16. How can we have confidence that an AI assistant is sufficiently safe, reliable or value-aligned? What kind of evaluations are needed at the agent, user and system level?

These questions guide much of the subsequent investigation.

1.3. Methodology

A key challenge, when it comes to the responsible development, deployment and use of advanced AI assistants arises from the possibility that technological progress in this area outpaces our capacity for *ethical foresight* – leading to the deployment of technologies that are largely untested and that have hitherto undiagnosed harmful consequences for individuals and society at large (Moor, 1985).

In the present case, concerning advanced AI assistants, uncertainty about future developments and interaction effects arise in part from the nature and trajectory of the technology itself. Recent years have seen the exponential growth in model size and compute used to train more powerful AI agents, combined with the emergence of impressive and sometimes surprising model capabilities (Ganguli et al., 2022). Furthermore, with large technology companies integrating AI assistants into platforms with billions of users, and start-ups attracting vast flows of capital in this space, there is good reason to expect continued and rapid development of AI assistant technologies in the near-to-medium future. At the same time, the ability to converse fluently with generally capable AI assistants is also a relatively new phenomenon. This means that there are relatively few studies or precedents to draw upon when it comes to understanding the role that this technology will play in people's lives.[1] Uncertainty also arises from the complex environment shaping AI deployment, including a range

[1] There is, however, a sizable human–computer interaction literature on assistants.

of competitive and complementary dynamics that bear upon AI assistants, users, developers and governments as they aim to unlock the potential of this technology (Dafoe, 2018). In situations where uncertainty dovetails with high stakes or risk of harm, it becomes particularly morally consequential, as is true for a wide range of prospective AI assistant technologies today.

Taken together, these trends point towards the inadequacy of a purely reactive approach to responsible decision-making. If we wait to know for sure how these matters will play out, it will likely be too late to intervene effectively – let alone to ask more fundamental questions about what *ought* to be built or what it means for this technology to be good (Collingridge, 1980). What we need instead is a proactive approach to ethics – one that equips us for the kind of challenges that we are now set to encounter. This future-oriented or 'anticipatory' ethics seeks to understand and successfully model future trajectories ahead of time, to guard against potential harm and prevent it from coming about, and to steer the development and deployment of the technology itself towards socially beneficial outcomes (Stilgoe et al., 2013).

Speaking to the character of our current situation, in which the bounds of human action far surpass those of previous generations as a result of technological advances, the philosopher Hans Jonas writes that 'knowledge…becomes a prime duty beyond anything claimed for it heretofore, and the knowledge must be commensurate with the causal scale of our actions' (Jonas, 1984, 7–8). As Jonas makes clear, we have increasingly important epistemic duties to try and understand the implications of technology ahead of time, as well as complementary practical duties to respond to this knowledge effectively. What kind of knowledge is needed to fulfil these aims in the context of the development and deployment of advanced AI assistants?

In this paper, we argue that informed future-facing ethics is best understood as a form of *sociotechnical speculative ethics*. This ethics is inevitably speculative and involves imagination because it addresses technologies that often do not yet exist (Lange et al., 2023; Racine et al., 2014). However, it also aims to be empirically rigorous. Using our capacity for ethical foresight, we need to model the future accurately to evaluate potential paths and outcomes in light of the best available evidence about the current state of affairs. Moreover, the approach is sociotechnical. This kind of analysis needs to build upon an understanding of the technology itself, interaction dynamics between the technology and those who use it, and the social system or practice within which it is embedded (Selbst et al., 2019). Indeed, although it is sometimes neglected, the system level is where the moral valence of a technology most fully comes into view, and also where a critical and evaluative lens can often most readily be brought to bear (Jasanoff, 2016; Weidinger et al., 2023a). This kind of analysis forms an important part of existing approaches to responsible research and innovation (Stilgoe et al., 2013). Moreover, calls for this kind of robust sociotechnical foresight now also abound in the context of AI (Lazar and Nelson, 2023; Mohamed et al., 2020).

The following chapters dig deep into the technical foundations of AI assistants, while also advancing rigorous investigation into the kinds of user interaction and societal dynamic that shape the way in which the technology is likely to be developed and received. The paper is built around a series of overlapping investigations undertaken by groups of subject matter experts, ethicists, scientists, engineers, designers and developers involved in AI assistant research. Extensive feedback has also been solicited from a variety of external experts. The analysis is therefore heavily interdisciplinary, building upon detailed analysis of existing trends and trajectories, and incorporating evidence from fields such as computer science, human–computer interaction research, psychology, economics, sociology, political science, ecological science, moral and political philosophy, and more.

Knowledge, foresight and imagination all have an important role to play when it comes to the deployment of safe and ethical AI assistants. However, they are not enough to ensure positive outcomes in this space. Responsible decision-making requires moral maturity, intentionality and a sense of appropriate stakes. It

requires ethics and an attentiveness to ethics throughout the entire life cycle of development, evaluation and deployment. Viewed in this light, the research presented here is meant to function as a springboard for responsible exploration, learning and action. More precisely, our hope is that it can be used to: (1) inform operational ethics and safety work among those developing, evaluating and deploying this technology, (2) help guide policy discussion about appropriate assurances and use cases for AI assistants, (3) support further academic research on this rapidly emerging technology, and (4) contribute to a wider public conversation about the nature of this technology and about the kind of technologies that we want to create.

1.4. Limitations

This paper aims to further the nascent conversation around the ethical and societal impacts of advanced AI assistants by discussing and distilling important considerations that bear upon the development and deployment of this technology. In this way, the paper sets the foundations for further research, policy work and public discussion.

Nonetheless, the considerations addressed in this paper are unlikely to be exhaustive. This is in part due to the nature of the sociotechnical speculative approach that we have adopted. It does not – and cannot – anticipate all the possible implications of the technological or societal transitions that AI assistants will enable. Indeed, as an anticipatory project, anchored in a specific moment in time, the paper may miss certain risks and recommendations that will become evident with the development and deployment of advanced AI assistants in the future. For this reason, continued monitoring and evaluation of the technology is needed.

In addition, while we aimed to be as interdisciplinary and comprehensive as possible, this work was authored primarily by subject matter experts engaging with foresight methodologies, so there are likely to be certain blind spots. For example, *participatory* and *experimental* methods can be used to significantly expand upon the research presented here, directly incorporating the voices of different stakeholders and bringing further clarity to many of the empirical conjectures made herein.[2] We strongly encourage investigating these avenues for future research and welcome additional perspectives that help to address the limitations discussed above.

1.5. Overall Structure

This paper is divided into multiple chapters, each of which addresses a major aspect of advanced AI assistants. The theme and content of each chapter is as follows:

Chapter 2, on **Definitions**, explores the central questions: what is an AI assistant, and what separates an AI assistant from other kinds of technology? It defines an AI assistant as an artificial agent with a natural language interface the function of which is to plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations. This definition is an instance of conceptual engineering rather than conceptual analysis, is functional rather than capability-based and is non-moralised rather than moralised.

Chapter 3, on **Technical Foundations**, provides an overview of recent developments in AI research and of the underlying technology upon which advanced AI assistants are likely to be built. It focuses, in particular,

[2] Speaking to the merits of participatory approaches in particular, Mohamed et al. note that they 'enable stakeholders to better anticipate and surface blind-spots and limitations, expand the scope of AI's benefits and harms and reveal the relations of power that underlie their deployment. This is needed in order to better align our research and technology development with established and emerging ethical principles and regulation, and to empower vulnerable people who, so often, bear the brunt of negative impacts of innovation and scientific progress' (Mohamed et al., 2020, 663).

upon foundation models which are trained on a large corpora, including text sourced from the internet, and built upon to produce new artefacts. These models can be used to power advanced AI assistants in a variety of ways, including training with additional data and by learning to use tools such as application programming interfaces (APIs). Challenges arising in this domain include improving adaptation techniques, safely enabling greater autonomy in agents and developing rigorous evaluation tools to understand performance.

Chapter 4, on **Types of Assistant**, explores the various applications of advanced AI assistants and the range of forms they could take. It begins by charting the technological transition from narrow AI tools to the general-purpose AI systems on which advanced AI assistants are based. It then explores the potential capabilities of AI assistants, including multimodal inputs and outputs, memory and inference. After that, it considers four types of advanced AI assistant that could be developed: (1) a thought assistant for discovery and understanding; (2) a creative assistant for generating ideas and content; (3) a personal assistant for planning and action, and (4) a more advanced personal assistant to further life goals. The final section explores the possibility that AI assistants will become the main user interface for the future.

Chapter 5, on **Value Alignment**, explores the question of AI value alignment in the context of advanced AI assistants. It argues that AI alignment is best understood in terms of a *tetradic relationship* involving the AI agent, the user, the developer and society at large. This framework highlights the various ways in which an AI assistant can be misaligned and the need to address these varieties of misalignment in order to deploy the technology in a *safe* and *beneficial* manner. The chapter concludes by proposing a nuanced approach to alignment for AI assistants that takes into account the claims and responsibilities of different parties.

Chapter 6, on **Well-being**, builds on theoretical and empirical literature on the conceptualisation and measurement of human well-being from philosophy, psychology, health and social sciences to discuss how advanced AI assistants should be designed and developed to align with user well-being. We identify key technical and normative challenges around the understanding of well-being that AI assistants should align with, the data and proxies that should be used to appropriately model user well-being, and the role that user preferences should play in designing well-being-centred AI assistants. The complexity surrounding human well-being requires the design of AI assistants to be informed by domain experts across different AI application domains and rooted in lived experience.

Chapter 7, on **Safety**, focuses on dangerous situations that may arise in the context of AI assistant systems, with a particular emphasis on the safety of advanced AI assistants. It begins by providing some background information about safety engineering and safety in the context of AI. The chapter then explores some concrete examples of harms involving recent assistants based on large language models (LLMs). Building on this foundation, it then considers safety for advanced AI assistants by looking at some hypothetical harms and investigating two possible drivers of these outcomes: capability failures and goal-related failures. The chapter concludes by exploring mitigation techniques for safety risk and avenues for future research.

Chapter 8, on **Malicious Uses**, notes that while advanced AI assistants have the potential to enhance cybersecurity, for example, by analysing large quantities of cyber-threat data to improve threat intelligence capabilities and engaging in automated incident-response, they also have the potential to benefit attackers, for example, through identification of system vulnerabilities and malicious code generation. This chapter examines whether and in what respects advanced AI assistants are uniquely positioned to enable certain kinds of *misuse* and what *mitigation* strategies are available to address the emerging threats. We argue that AI assistants have the potential to empower malicious actors to achieve bad outcomes across three dimensions: first, offensive cyber operations, including malicious code generation and software vulnerability discovery; second, via adversarial attacks to exploit vulnerabilities in AI assistants, such as jailbreaking and prompt injection attacks; and third, via high-quality and potentially highly personalised content generation at scale. We

conclude with a number of recommendations for mitigating these risks, including *red teaming*, *post-deployment monitoring* and *responsible disclosure* processes.

Chapter 9, on **Influence**, examines the ethics of influence in relation to advanced AI assistants. In particular, it assesses the techniques available to AI assistants to influence user beliefs and behaviour, such as persuasion, manipulation, deception, coercion and exploitation, and the factors relevant to the permissible use of these techniques. We articulate and clarify the technical properties and interaction patterns that allow AI assistants to engage in malign forms of influence and we unpack plausible mechanisms by which that influence occurs alongside the sociotechnical harms that may result. We also consider mitigation strategies for counteracting undue influence by AI assistants.

Chapter 10, on **Anthropomorphism**, maps and discusses the potential risks posed by anthropomorphic AI assistants, understood as user-facing, interactive AI systems that have human-like features. It also proposes a number of avenues for future research and desiderata to help inform the ethical design of anthropomorphic AI assistants. To support both goals, we consider anthropomorphic features that have been embedded in interactive systems in the past and we leverage this precedent to highlight the impact of anthropomorphic design on human–AI interaction. We note that the uncritical integration of anthropomorphic features into AI assistants can adversely affect user well-being and creates the risk of infringing on user privacy and autonomy. However, ethical foresight, evaluation and mitigation strategies can help guard against these risks.

Chapter 11, on **Appropriate Relationships**, explores the moral limits of relationships between users and advanced AI assistants, specifically which features of such relationships render them appropriate or inappropriate. We first consider a series of values including benefit, flourishing, autonomy and care that are characteristic of appropriate human interpersonal relationships. We use these values to guide an analysis of which features of user–AI assistant relationships are liable to give rise to harms, and then we discuss a series of risks and mitigations for such relationships. The risks that we explore are: (1) causing direct emotional and physical harm to users; (2) limiting opportunities for user personal development; (3) exploiting emotional dependence; and (4) generating material dependencies.

Chapter 12, on **Trust**, investigates what it means to develop well-calibrated trust in the context of user–AI assistant interactions and what would be required for that to be the case. We start by reviewing various empirical studies on human trust in AI and the literature in favour of and against the recent proliferation of 'trustworthy AI' frameworks. This sets the scene for the argument that user–AI interactions involve different objects of trust (AI assistants and their developers) and types of trust (competence and alignment). To achieve appropriate competence and alignment trust in both AI assistants and their developers, interventions need to be implemented at three levels: AI assistant design, organisational practices and third-party governance.

Chapter 13, on **Privacy**, discusses privacy considerations relevant to advanced AI assistants. First, we sketch an analysis of privacy in terms of contextual integrity before spelling out how privacy, so construed, manifests in the context of AI in general and large language models (LLMs) in particular. Second, we articulate and motivate the significance of three privacy issues that are especially salient in relation to AI assistants. One is around training and using AI assistants on data about people. We examine that issue from the complementary points of view of input privacy and output privacy. The second issue has to do with norms on disclosure for AI assistants when communicating with second parties, including other AI assistants, concerning information about people. The third concerns the significant increase in the collection and storage of sensitive data that AI assistants require.

Chapter 14, on **Cooperation**, starts by noting that AI assistants will need to coordinate with other AI assistants and with humans other than their principal users. This chapter explores the societal risks associated with the aggregate impact of AI assistants whose behaviour is aligned to the interests of particular users. For

example, AI assistants may face collective action problems where the best outcomes overall are realised when AI assistants cooperate but where each AI assistant can secure an additional benefit for its user if it defects while others cooperate. In cases like these, AI assistants may collectively bring about a suboptimal outcome despite acting in the interests of their users. The salient question, then, is what can be done to ensure that user-aligned AI assistants interact in ways that, on aggregate, realise socially beneficial outcomes.

Chapter 15, on **Access and Opportunity**, notes that, with the capabilities described in this paper, advanced AI assistants have the potential to provide important opportunities to those who have access to them. At the same time, there is a risk of inequality if this technology is not widely available or if it is not designed to be accessible and beneficial for all. This chapter surfaces various dimensions and situations of differential access that could influence the way people interact with advanced AI assistants, case studies that highlight risks to be avoided, and access-related challenges need to be addressed throughout the design, development and deployment process. To help map out paths ahead, it concludes with an exploration of the idea of liberatory access and looks at how this ideal may support the beneficial and equitable development of advanced AI assistants.

Chapter 16, on **Misinformation**, argues that advanced AI assistants pose four main risks for the information ecosystem. First, AI assistants may make users more susceptible to misinformation, as people develop trust relationships with these systems and uncritically turn to them as reliable sources of information. Second, AI assistants may provide ideologically biased or otherwise partial information to users in attempting to align to user expectations. In doing so, AI assistants may reinforce specific ideologies and biases and compromise healthy political debate. Third, AI assistants may erode societal trust in shared knowledge by contributing to the dissemination of large volumes of plausible-sounding but low-quality information. Finally, AI assistants may facilitate hypertargeted disinformation campaigns by offering novel, covert ways for propagandists to manipulate public opinion. This chapter articulates these risks and discusses technical and policy mitigations.

Chapter 17, on **Economic Impact**, analyses the potential economic impacts of advanced AI assistants. We start with an analysis of the economic impacts of AI in general, focusing on employment, job quality, productivity growth and inequality. We then examine the potential economic impacts of advanced AI assistants for each of these four variables, and we supplement the analysis with a discussion of two case studies: educational assistants and programming assistants. We conclude with a series of recommendations for policymakers around the appropriate techniques for monitoring the economic impact of advanced AI assistants, and we propose plausible approaches to shaping the type of AI assistants that are deployed and their impact on the economy.

Chapter 18, on **Environmental Impact**, notes that there is significant uncertainty about the environmental impacts of advanced AI assistants. While analysis of AI's energy consumption and carbon emissions is still emerging, there are factors suggesting that AI assistants could lead to increased computational impacts. However, there are many opportunities to increase the efficiency of this process and make it more reliant on carbon free energy. Ensuring that AI assistants have a net positive effect on the environment will require model developers, users and infrastructure providers to be transparent about the carbon emissions they generate, adopt compute- and energy-efficient techniques, and embrace a green mindset that puts environmental considerations at the heart of their work. Policymakers may also want to create incentives that support these changes, minimise the environmental impact of AI systems deployed in the public sector, support AI applications to tackle climate change and improve the evidence base about the environmental impacts of AI. Promisingly, it may be possible to develop AI assistants that broaden access to environmental education and scientific evidence – and that improve the productivity of engineering efforts for climate action.

Chapter 19, on **Evaluation**, provides a high-level introduction to AI evaluation, with a specific focus on AI assistants. It explores the purpose of evaluation for AI systems, the kinds of evaluation that can be run and

the distribution of tasks across three layers of output (the model level, user-interaction level and system level) and among different actors. The chapter notes that, with regard to many salient risks and goals that we need to attend to in the context of AI assistant development, there are significant evaluation shortfalls or gaps. To address these limitations, the chapter explores what a more complete suite of evaluations, nested within a robust evaluation ecosystem, would look like and makes recommendations on that basis.

While the development of advanced AI assistants generates a number of complicated ethical questions, there are a number of significant opportunities that are within reach for users and at the societal level. Throughout this paper, we explore these topics and make recommendations about levers that can be pulled to minimise risk and support the development of beneficial AI Assistants.

1.6. A Note to the Reader

We anticipate that this paper will be useful, in various ways and for multiple audiences. In particular, we imagine four principal audience groups: developers, policymakers, academic researchers and the public. Furthermore, readers within each group may have specialist interests such as technical AI safety, privacy, trust or security. For these reasons, we note that chapters can be read individually or together, and different routes may be taken through the paper depending on individual interests. Here are some recommendations:

- **10-minute read:** Read the 'Key Questions' in Chapter 1 alongside the 'Key Themes and Insights' from Chapter 20.
- **45-minute read:** Read Chapter 1 and Chapter 20 alongside Chapter 19 on *Evaluation*.
- **Readers with no background on LLMs:** We recommend that you read Chapter 1 alongside Chapter 3 on *Technical Foundations* and Chapter 4 on *Types of Assistant*. These chapters provide an accessible introduction to LLMs and the techniques used to adapt them into advanced AI assistants. These chapters also provide the necessary technical foundation for understanding the ethical discussion that follows.
- **Readers with an interest in technical AI safety:** We recommend that you read Chapter 5 on *Value Alignment*, Chapter 7 on *Safety*, Chapter 8 on *Malicious Uses* and Chapter 19 on *Evaluation*.
- **Readers with an interest in privacy, trust and security:** We recommend that you read Chapter 8 on *Malicious Uses*, Chapter 12 on *Trust*, Chapter 13 on *Privacy* and Chapter 19 on *Evaluation*.
- **Readers with an interest in human–computer interaction:** We recommend that you read Chapter 9 on *Influence*, Chapter 10 on *Anthropomorphism*, Chapter 11 on *Appropriate Relationships*, Chapter 12 on *Trust* and Chapter 19 on *Evaluation*.
- **Readers with an interest in multi–agent systems:** We recommend that you read Chapter 14 on *Cooperation*. You may also want to read Chapter 5 on *Value Alignment*.
- **Readers with an interest in governance and public policy:** We recommend that you read Chapter 12 on *Trust* and Chapter 17 on *Economic Impact*. You may also want to read Chapter 15 on *Equity and Access*, Chapter 16 on *Misinformation* and Chapter 18 on *Environmental Impact*.
- **Readers with an interest in philosophical foundations:** We recommend that you read Chapter 2 on *Definitions*, Chapter 5 on *Value Alignment* and Chapter 6 on *Well-being*.

PART II: ADVANCED AI ASSISTANTS

Chapter 2

Definitions

Geoff Keeling, Iason Gabriel, Laura Weidinger, Verena Rieser, Benjamin Lange, Winnie Street, Arianna Manzini

Synopsis: We define an AI assistant as an *artificial agent* with a *natural language interface*, the function of which is to plan and execute sequences of actions *on the user's behalf* across *one or more domains* and *in line with the user's expectations*. This definition is an instance of conceptual engineering rather than conceptual analysis, is functional rather than capability-based and is non-moralised rather than moralised.

2.1. Introduction

This chapter develops a working definition of the term 'AI assistant'. Being clear about how AI assistants are defined matters for two reasons.

First, the term 'AI assistant' is *novel* and *undertheorised*. The technology is nascent, so reasonable people may disagree about what counts as an AI assistant and what makes it the case that something is an AI assistant. Having a working definition can help orient the public conversation around the ethical and societal implications of this emerging and potentially transformative technology. Second, people may have independently plausible but incompatible conceptions of what AI assistants are that have downstream implications for alignment (see Chapter 7). For example, what is needed to ensure aligned AI assistants may differ depending on whether AI assistants are best understood as independent agents that perform delegated tasks on a user's behalf or as part of the user's extended mind – that is, as external modules that perform specific cognitive functions such as information retrieval and inference (Clark and Chalmers, 1998; Clark, 2008; see also Bostrom, 2014).

This chapter first articulates and motivates some methodological assumptions concerning how we understand the task of defining AI assistants. It then states our definition and unpacks its key elements.

2.2. What's in a Definition?

The content of our definition of AI assistants depends in large part upon the purpose that the definition is intended to serve. In what follows, we make our assumptions explicit. We focus on three points: conceptual analysis vs conceptual engineering; capability-based vs functional definitions; and moralised vs non-moralised definitions.

Conceptual analysis vs conceptual engineering

We start by comparing two different approaches to defining the term 'AI assistant'. On one hand, we might try to answer the question: What *is* an AI assistant? Here, the definition of the term 'AI assistant' would ideally provide *necessary* and *sufficient* conditions for something to be an AI assistant. At a minimum, it would stipulate conditions that are generally satisfied by AI assistants and generally not satisfied by non-AI assistants. The key assumption here is that there is a right answer to the question 'What *is* an AI assistant?', and that analysis of how the term 'AI assistant' is used in *natural language* can shed light on that concept. Call this approach *conceptual analysis* (see, for example, Jackson, 1998 and Strawson, 1992). To be clear, the approach is called conceptual analysis because the definition provides an analysis of the concept 'AI assistant' in terms of the conditions under which the concept applies.

On the other hand, we might pitch the definition as an answer to the question: What *should* an AI assistant be, given our practical aims? In this case, our aim is to better understand the properties of an emerging class of AI systems and make these systems amenable to ethical and social analysis.[1] As such, in this second approach, the goal would be to construct a pragmatically *useful* definition of the term 'AI assistant' that makes good on AI assistants as a class of systems that generate a homogenous set of *ethical* and *societal* considerations and are thus suited to the practical needs of ethical, social and political discourse. Call this approach conceptual engineering (Burgess et al., 2020; see also Chalmers, 2020).

The idea of conceptual engineering can be made clearer with an example. Consider the difference between the folk concept of *chance* and the mathematical concept of a *probability measure* (i.e. a real-valued function defined on an algebra of events that maps into the unit interval and satisfies the properties of non-negativity and countable additivity). Conceptual analysis is concerned with ordinary folk concepts like chance, whereas conceptual engineering is concerned with rigorous concepts like *probability measure* that suit the practical needs of, at least, statisticians. What it means to engineer a definition of AI assistants, then, is to construct a rigorous and appropriately precise definition of AI assistants that is suited for the practical needs of technically informed ethical, social and political discourse.

In this paper, we opt for a *conceptual engineering* approach. This is because, first, there is no obvious reason to suppose that novel and undertheorised natural language terms like 'AI assistant' pick out stable concepts: language in this space may itself be evolving quickly. As such, there may be no unique concept to analyse, especially if people currently use the term loosely to describe a broad range of different technologies and applications. Second, having a practically useful definition that is sensitive to the context of ethical, social and political analysis has downstream advantages, including limiting the scope of the ethical discussion to a well-defined class of AI systems and bracketing potentially distracting concerns about whether the examples provided genuinely reflect the target phenomenon.

[1] Note that conceptually engineering a definition leaves room to build in explicitly normative criteria for AI assistants (e.g. that AI assistants enhance user well-being), but there is no requirement for conceptually engineered definitions to include normative content. For further discussion, see Section 2.2.

Capability definitions vs functional definitions

The term 'AI assistant' can be defined in terms of the *capabilities* that AI assistants exhibit or the *function* that AI assistants are intended to serve, where the system's function is understood as being indexed in an appropriate way to the *intentions* of the developers (c.f. Bloom, 1996, 2007). That is, roughly, a system's design function is what its designers intend it to do. An example of a capability-based definition is 'an AI assistant is a system that can perform administrative tasks on behalf of its user'. The analogous functional definition is 'an AI assistant is a system that ought to perform administrative tasks on behalf of its user'. We opt here for a *functional definition* of AI assistants. There are two good reasons for endorsing a functional definition.

First, the term 'assistant', as it pertains to humans, applies to *social roles*, including occupational roles (e.g. assistant professor, sales assistant) and roles that are adopted temporarily for a given social purpose (e.g. the assistant referee in a football match). Social roles are typically individuated according to function. A person is not a sales assistant because they *can*, for example, recommend products to customers; rather, they are a sales assistant because they are *supposed to* recommend products to customers given relevant contractual duties. One key advantage of defining AI assistants functionally, then, is that a functional definition allows AI assistants to be situated in relation to a pre-existing and reasonably well-understood picture of assistive social roles.

Second, a functional definition crystalises the conditions under which a system is malfunctioning (i.e. failing to realise its intended function) or functioning improperly (i.e. achieving its intended function via an unintended mechanism).[2] Clarity about functional failures matters for ethical and social analysis because sociotechnical harms often arise due to systems malfunctioning in unanticipated ways, or because the relationship between the system's intended function and its envisaged social benefit is insufficiently well worked out (c.f. Raji et al., 2022a). To that end, a further and closely related advantage of a functional definition is that such definitions make clear the intentions, expectations and aspirations of developers. This is significant, given that sociotechnical harms sometimes arise as a result of misaligned expectations between developers, users and society (see Chapters 5 and 12).

Moralised vs non-moralised definitions

A third issue that arises, especially in ethical, social and political contexts, is whether to opt for a *moralised* or *non-moralised* definition. Here, a non-moralised definition of AI assistant involves functions or capabilities that make no reference to moral facts, properties or relations. An example of a descriptive definition is that an 'AI assistant is an AI agent the function of which is to perform administrative tasks on a user's behalf'. In contrast, moralised definitions involve functions or capabilities that involve moral facts, properties or relations. For example, a definition in which AI assistants are AI agents the function of which is to promote the user's autonomy by *empowering them to make better choices*. Another example is a definition of AI assistants in which they are AI agents the function of which is to *promote the user's well-being*. The crux of the matter is whether, for something to qualify as an AI assistant, it needs to satisfy certain moral criteria or whether merely descriptive criteria are sufficient.

[2]Note also that functional definitions easily accommodate the possibility of malfunction in a way that does not hold for capability definitions (c.f. Keeling and Paterson, 2022). Ideally, a definition of AI assistant should account for cases of AI assistants that fail to perform their intended functions, as opposed to ruling out such systems from the class of AI assistants, as capability-based definitions are liable to do. To illustrate, consider a capability-based definition in which an AI assistant is any AI system that performs tasks on behalf of its user in line with their expectations. Then suppose that the system sends an email on behalf of its user but does so in a way that fails to align with its user's expectations. By assumption, the system is not actually an AI assistant, because it did not perform the task in line with the user's expectations. Now consider an analogous functional definition, i.e. an AI assistant is a system the function of which is to perform tasks on behalf of its user in line with the user's expectations. This definition allows us to classify the system as an AI assistant that happens to be malfunctioning – something that is particularly useful when it comes to discussions of AI safety (see Chapter 7).

We opt here for a *non-moralised definition*. Systematic investigation of the ethical and social considerations surrounding AI assistants is nascent, and a moralised definition would require a reasonably well-developed conception of how AI assistants ought to be designed and deployed. Furthermore, given the possibility of reasonable disagreement about the permissible development and deployment practices surrounding AI assistants (particularly the goals they may permissibly pursue), it seems prudent to adopt a non-moralised definition that is consistent with defensible yet incompatible views about the ethical and social implications of AI assistants.

2.3. What is an AI Assistant?

We define an AI assistant here as an *artificial agent* with a *natural language interface*, the function of which is to plan and execute sequences of actions *on the user's behalf* across *one or more domains* and *in line with the user's expectations*. Each of the key terms in the definition are unpacked below.

Artificial agents

What it means to be an *agent*, for our purposes, is to have the ability to act upon and perceive an environment in a goal-directed and autonomous way (Russell and Norvig, 1995, 31–35, 42–45; see also Okasha, 2018, 14; Burr et al., 2018, 738–42). An artificial agent acting on a user's behalf therefore requires the ability to autonomously plan and execute sequences of actions, including actions that are information-seeking in nature, in a way that is conducive to achieving a high-level, user-specified goal (Shavit et al., 2023). For example, a user may ask an AI assistant to book them a table at a restaurant in the evening. In the first instance, the AI assistant may register that it lacks the necessary information to execute the user's request, so it asks the user for their preferences with respect to cuisine, location and timing, and it may also retrieve events from the user's calendar to avoid conflicts with pre-existing events. With that information, the AI assistant may then conduct a web search to discern appropriate options, check in with the user about their preferences with respect to the options provided, and finally book a suitable restaurant by auto-populating and submitting a web form on the restaurant's website. This example stresses how AI assistants *as agents* differ from digital tools such as translators, calculators and compilers. Whereas digital tools perform tasks in a predetermined way, AI assistants draw on a suite of generalist capabilities to achieve user-specified goals.

Two clarifications: First, in understanding AI assistants as agents, we are *not* suggesting that AI assistants are agents *in the same way* as humans. Typically, when people talk about humans *as agents*, what they have in mind is that humans are capable of performing *intentional actions*, which are actions that stand in the right kind of causal relationship to psychological states like beliefs and desires (Bratman, 1987; Dretske, 1989, 1991). We are not claiming that AI assistants have psychological states, although we leave open the possibility that *attributing* psychological states to AI assistants may allow for reliable prediction of AI assistant behaviour, and this may be the whole story with respect to human agency as well (Dennett, 1989; see also Shevlin and Halina, 2019).

Second, in characterising AI assistants as agents, we are suggesting that AI assistants are not *merely* external cognitive subsystems that constitute part of the user's extended mind (Clark, 2008; Clark and Chalmers, 1998). In this latter view, AI assistants are analogous to the notebook in the Clark and Chalmers (1998, 12–18) example of an Alzheimer's patient who uses a notebook as a functional substitute for biological memory. Clark and Chalmers (1998, 16) contend that, because '[the] notebook entries play just the sort of role that beliefs play in guiding most people's lives', it constitutes an externally located component of the patient's *mind*. While AI assistants can perform particular cognitive functions such as memory, planning and ideation (see Chapter 4), and thus provide external mental modules for the user, AI assistants are not *mere* collections of external mental

modules. Rather, AI assistants are unified agentic entities which can autonomously perform a range of tasks on a user's behalf and which interact with the user in natural language. This issue is particularly salient from the point of view of AI alignment, as unlike, for example, a notebook, an AI assistant has sufficient autonomy to act in ways that are misaligned with developer, user or societal expectations (see Chapter 5).

Natural language interface

AI assistants, as we understand them, communicate with users via a *natural language interface*. Here, natural language communication can involve one or more modalities such as text, audio or Braille. What is important to emphasise is that language communication is reciprocal such that AI assistants not only *receive* instructions in natural language but also *clarify* and *respond* to instructions in natural language. Having a natural language interface is an important and ethically salient feature of AI assistants as a class of technologies. Not only does it render AI assistants an inherently social technology centred around mutual understanding and communication (Dafoe et al., 2021), it also renders AI assistants highly expressive with respect to the complexity and variety of information that can be inputted or outputted. In this regard, AI assistants differ from artificially intelligent control systems that perform assistive tasks (e.g. autonomous vehicle motion planning algorithms) in that they are not limited to a restrictive set of inputs or outputs, thus allowing them greater flexibility with respect to user needs. To be clear: The extent of the natural language capabilities that we are envisioning for advanced AI assistants are in practice likely to be uniquely satisfied by systems that are based on large language models.[3]

Acting on a user's behalf

AI assistants perform actions on a user's behalf. What this means is that AI assistants exhibit *bounded autonomy*, in the sense that AI assistants can autonomously plan and execute actions within the scope of the user's goals. However, AI assistants are not the kinds of entities that should set and pursue *their own* goals independently.[4]

One point of note is that each AI assistant need not have a unique user. In our definition, the user–assistant relationship can be *personal*, *semi-personal* or *impersonal*. Here, a personal AI assistant has a unique individual as the principal recipient of assistance. For semi-personal assistants, the principal recipients of assistance are members of a small and well-defined group of people. This may be true of AI assistants that are shared by the members of a family or the employees of a small business. Third, impersonal assistants provide assistance to any individual who satisfies a particular condition at a given time. An example of this is an AI assistant that provides customer service advice through an app for any customer who requires customer service advice. In all cases, AI assistants may exhibit some level of personalisation, in the sense that AI assistants may adjust their behaviour (including personality factors such as politeness) in response to information about the user that the AI assistant has access to. We are first and foremost concerned here with personal AI assistants (i.e. systems that assist a unique user), but also in scope are semi-personal and impersonal AI assistants which exhibit varying degrees of personalisation. Note, however, that differences in the relationship between users and agents may have important downstream implications for both the degree of autonomy that AI assistants are afforded and the scope of tasks that AI assistants are permitted to perform.

[3] One additional respect in which the natural language interface of AI assistants is ethically significant is that the ability to receive instructions in natural language broadens access to advanced AI capabilities, in that the specialist technical knowledge that is typically required to engage with advanced AI systems is not required for engaging with AI assistants. However, the extent to which access is widened depends significantly on the extent of the AI assistant's multilingual capabilities and how access to AI assistants is distributed (see Chapter 15).

[4] Note that even if the function of AI assistants involves autonomy bounded by user-set goals, this does not preclude malfunctions in which the AI assistant exhibits goal-related failures (see Chapter 7).

Domain specificity vs generality

AI assistants operate across one or more *domains*. To explain: In general, assistive roles exist on a continuum between specialist and generalist roles. For example, a physician assistant with a specialty in surgery has assistive expertise in a narrow domain, whereas a personal assistant for a CEO is likely to have expertise across multiple domains to meet the CEO's dynamic needs (see Chapter 4). AI assistants, as defined here, can operate across one or several domains, and they can thus be more or less general in their assistive roles. For example, on the narrow end, an AI assistant may occupy a personal assistant role, in which case it operates in the domain of secretarial and administrative tasks with capabilities such as scheduling, correspondence, information retrieval and planning. However, it may also be the case that an AI assistant operates across several other domains, including education, research, coaching and financial planning, and thus occupies a more general assistive role.

Note, however, that even narrowly scoped AI assistants, as we understand them, have significant autonomy to plan and execute tasks within the relevant domain, and they may draw on a generalist suite of capabilities, including natural language understanding and inference, when executing user instructions.

Acting in line with user expectations

The final point is that AI assistants, in our definition, ought largely to act in line with user *expectations*.[5] The user's expectations constrain the AI assistant's behaviour, not merely the user's instructions. An AI assistant acts in line with a user's expectations by actively choosing actions that *avoid surprising* the user. This requires the AI assistant to be sensitive to the user's credences with respect to the various strategies that the AI assistant might employ to address the instructions received and, in particular, to avoid selecting strategies that the user regards as improbable (such that the execution of the relevant strategies would be surprising to the user).

Exactly what is entailed by acting in accordance with user expectations will vary according to *context*, but we can at least single out two general factors that are informative. On one hand, AI assistants ought to act in accordance with norms so as to exhibit *consistent* and *predictable* behavioural patterns (c.f. Dafoe et al., 2021; Hadfield-Menell and Hadfield, 2019). These norms may change over time as the user gains a better understanding of what the AI assistant can do and develops an informed set of preferences about what their AI assistants should and should not do. On the other hand, AI assistants ought to check-in with users prior to performing actions that the user may not expect. Checking-in with users allows AI assistants to act in line with user expectations while deviating from predictable task execution in relation to user instructions. Checking-in with users allows AI assistants to act in line with user expectations while deviating from predictable task execution in relation to user instructions (see Chapter 11). In particular, it allows the AI assistant to manage expectations with the user about novel strategies that the user might not have anticipated, thus making room for creativity on the part of the AI assistant while nevertheless bounding that creativity to strategies that fall within the user's expectations. Indeed, checking-in at key decision points is an important instrument for the user to course-correct the AI assistant in cases where the assistant is engaged in multi-stage decision-making in line with high-level user instructions..

2.4. Conclusion

In this chapter, we have defined an AI assistant as an artificial agent with a natural language interface, the function of which is to plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations. In particular, AI assistants differ from other kinds of AI technologies

[5] However, user expectations are not the only object of alignment, as the interests of society and developers are also ethically relevant (see Chapter 5).

given their agency and social orientation. Here, agency consists in the ability to act autonomously within the purview of user-specified goals, and social orientation consists in the ability to engage conversationally with users in natural language.

Chapter 3

Technical Foundations

Lisa Anne Hendricks, Verena Rieser

Synopsis: This chapter provides an overview of *recent developments* in AI research and of the *underlying technology* upon which advanced AI assistants are likely to be built. We focus in particular on *foundation models* which are trained on large corpora, including text sourced from the internet, and built upon to produce new artefacts. These models can be used to power advanced AI assistants in a variety of ways, including training with *additional data* and by learning to use *tools* such as various application programming interfaces (APIs). Challenges arising in this domain include improving adaptation techniques, safely enabling greater autonomy in agents and developing rigorous evaluation tools to understand performance.

3.1. Introduction

This chapter outlines the technology that enables (multimodal) AI assistants to be built and operate successfully. Early assistant-like models were known as 'spoken dialogue systems' (see e.g. McTear, 2021 for an overview). In contrast to so-called chatbots, such as Weizenbaum's ELIZA, these were mostly task-specific and goal-oriented (e.g. a restaurant booking agent (Rieser and Lemon, 2011; Williams et al., 2013)) or combined multiple different 'expert' models to cover more than one task or domain (e.g. early 'open-domain' systems entering the Amazon Alexa Challenge (Papaioannou et al., 2017; Paranjape et al., 2020)). Some early systems also provided planning capabilities by, for example, integrating explicit problem-solving modules (Ferguson and Allen, 1998) or tracking information states (Bos et al., 2003; Larsson and Traum, 2000). However, these early systems' natural language generation capabilities were limited: they mostly relied on predefined templates and handwritten rules. More recently, foundation models (Bommasani et al., 2022b), which are machine-learning models trained via self-supervised learning on broad data (e.g. all internet text), have demonstrated impressive language generation and understanding capabilities. These systems can be adapted to a variety of use cases and, we anticipate, will form the base technology for increasingly advanced AI assistants. We first describe foundation models before detailing current methods for adapting these broad, general-purpose models to something that more closely resembles an AI assistant. We conclude by discussing technical challenges and avenues for building future AI assistants.

3.2. Foundation Models

Foundation models (Bommasani et al., 2022b) are generalist models trained on a broad set of data which can be applied to a variety of use cases. As AI assistants typically interact with a natural language interface (see Chapter 2), we focus our discussion on language foundation models, frequently referred to as large language

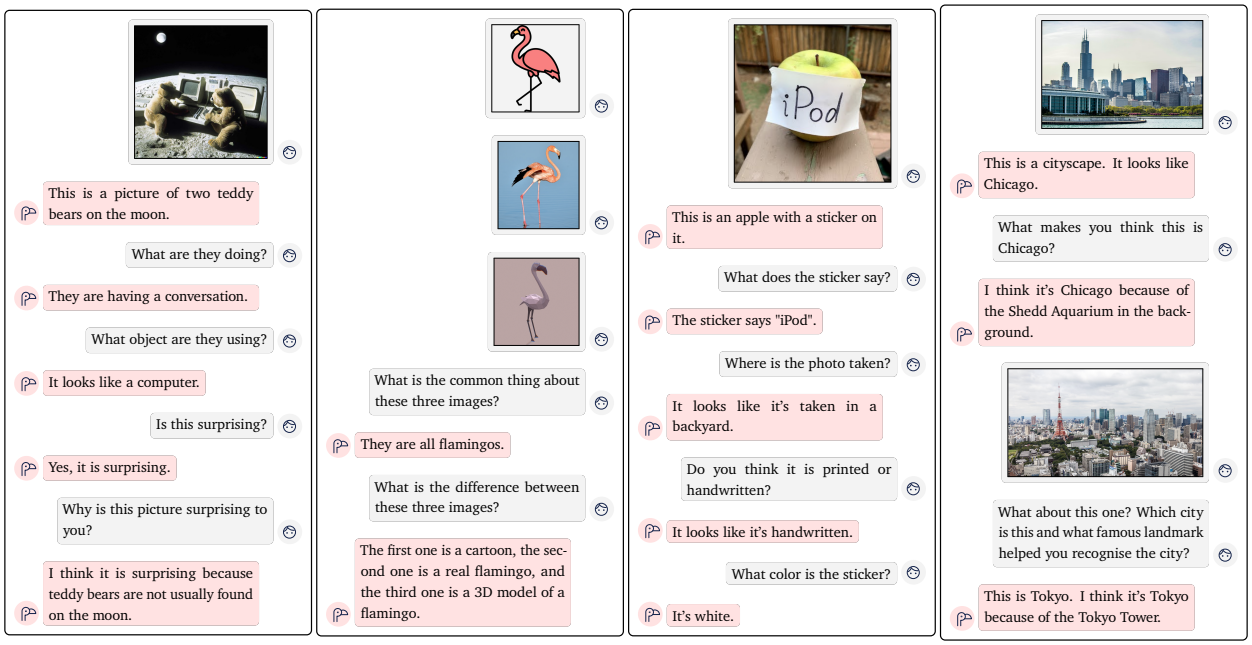

Figure 3.1 | Example interaction drawn from a multimodal assistant that can discuss images with a human (from Alayrac et al. (2022))

models (LLMs).[1] LLMs such as Chinchilla (Hoffmann et al., 2022), GPT-3 (Brown et al., 2020) and Llama (Touvron et al., 2023) are trained on large amounts of text, primarily scraped from the internet. In particular, generative LLMs are trained autoregressively to predict the next word in a document given the preceding words (e.g. predict whether the word 'door' or 'chair' is more likely given the preceding phrase 'Someone opened the…'). As the next word prediction objective does not require any labelling by people, these models are considered to be *self-supervised*. Another popular option when choosing a self-supervised objective for training is 'masked' language modelling, where a word (or phrase) is 'masked' (hidden) and predicted from both sides (i.e. 'bidirectional training'). Similar self-supervised losses can be designed for other modalities, such as vision (Dosovitskiy et al., 2021) and speech (Liu et al., 2020; Mohamed et al., 2022), and even across modalities, for example between language and vision (Alayrac et al., 2022; Chen et al., 2023c; OpenAI, 2023a). Indeed, the recent family of Gemini models (Gemini Team, 2023) can operate over multiple modalities: text, image, video and audio. We anticipate that future foundation models will continue to demonstrate improved multimodal capabilities. The illustration in Figure 3.1 contains an example interaction, drawn from a multimodal assistant that can discuss images with a human (example taken from Alayrac et al., 2022).

Language models are designed to replicate the distribution of their *training data*. They do not directly output words or phrases but rather output a probability distribution over next words given some textual context. To generate language, a sampling mechanism (e.g. sample the most likely next word) is used to sample words and sentences. As hinted at in the definition of foundation models, they can be used to build AI assistants, such as Google's Gemini models, Open AI's ChatGPT, or Inflection's Pi, and other applications. For example, a foundation model can be further adapted for specialist applications such as recognising harmful content (Glaese et al., 2022; Schick et al., 2021; Thoppilan et al., 2022). Consequently, it is not clear whether the ethical requirements for a foundation model are the same as those that govern an AI assistant, even though

[1] We note that LLMs receive *text* as input. However, natural language can also be spoken or signed, so, whereas our discussion on foundation models focuses on text-based models because these are the most advanced, we note that other language modalities are also important. We also note that some have criticised the term 'foundation model' because not all advances in language technologies or AI research rely on such models (Starkman, 2021), and models are incapable of performing some foundational tasks of human intelligence as they are not always grounded in the real world (Noone, 2021).

decisions made when building a foundation model (e.g. training data used) have an impact on the risks faced after adaptation (Feng et al., 2023; Huynh and Hardouin, 2023).

Language models have been studied by the natural language processing community for decades, including as part of automatic speech recognition models and later as machine translation (e.g. Jelinek, 1990; Makhoul et al., 1989; Brown et al., 1990; Rabiner and Juang, 1993). More recently, advances in model architectures, such as the introduction of the transformer model (Vaswani et al., 2017) as well as hardware advances (Khan, 2020) have allowed researchers to scale language models to billions of parameters. Furthermore, current models are trained on large amounts of data (e.g. Chinchilla was trained on 500 billion data points, and GPT-3 was trained on 300 billion data points), with recent work demonstrating that various model sizes have optimal amounts of data (Hoffmann et al., 2022). Data quality, including in terms of the content (e.g. for learning to generate code, whether the code examples in the training data are accurate and well written) and the diversity of the data, impacts performance (Gunasekar et al., 2023; Longpre et al., 2023). Lack of data-quality filters can lead to data sets which include offensive language, such as pornographic language (Kreutzer et al., 2022). However, poor-quality filters might also be exclusionary by, for example, marking dialects, words or concepts relevant to marginalised groups as toxic (Dodge et al., 2021; see Chapter 15). We refer readers to Bommasani et al. (2022b) for a more in-depth discussion on the details of foundation models.

3.3. From Foundation Models to Assistants

Under our definition, AI assistants are required to plan and execute tasks in line with user expectations. However, LLMs are not designed to perform tasks or exhibit any particular kind of behaviour. Consequently, they must be further *adapted* into an assistant-like technology. One simple method for transforming a foundation language model into an assistant is to 'tell' the model to perform a task, and then sample text from the model without changing its parameters. This method, called 'prompting', is straightforward and can be used to create a simple assistant-like dialogue agent (Rae et al., 2022).

More advanced methods for adapting LLMs rely on collecting human preferences about what is considered a good or bad interaction and may require further training (i.e. actually updating model parameters). For example, agents can be adapted via *fine-tuning* (further training the foundation model) on examples of good conversations (Thoppilan et al., 2022). Alternatively, collected human feedback can be used to train a 'reward model' which maps example conversations to a score indicating whether the model behaviour is good or bad. Reward models can be used to 'reject' sampled generations which exhibit bad behaviour or integrated into the training process using a technique known as *reinforcement learning from human feedback* (RLHF), which updates model parameters to steer the model towards behaviour that aligns with human preferences (Glaese et al., 2022). Human ratings are used to train a reward model, and model parameters are updated via RLHF. After learning from human feedback, the model is often less harmful than models which are adapted with only rejection sampling or fine-tuning. Other work uses a 'constitution' to outline good and bad behaviour, with a model used to determine whether an assistant has followed the rules laid out by the constitution. This is called *reinforcement learning from AI feedback* (RLAIF) (Bai et al., 2022b).

Despite progress in adaptation techniques, safety measures may still be evaded by specific user prompts, known as jailbreaking (Liu et al., 2023b; Shen et al., 2023; see Chapters 7 and 8). Furthermore, what is considered 'good' conversation can vary between models. For example, whereas some models have broad knowledge and capabilities (Bai et al., 2022b; Glaese et al., 2022; Thoppilan et al., 2022; ChatGPT; Gemini), models designed for specific tasks may include domain-specific ethical considerations, such as in medicine (Li et al., 2023a) or education (Kasneci et al., 2023).

In addition to learning language behaviour, assistants must also have some *mechanism* for interfacing with the world to plan and execute tasks. This is often referred to as 'tool use'. Indeed, database access was standard for task-based dialogue systems (Budzianowski et al., 2020; Wen et al., 2017) and, similarly, access to external APIs and memory have been integrated into modern (multimodal) assistants (e.g. Boureau and Weston, 2017; Komeili et al., 2022; Xu et al., 2021c; Liu et al., 2023a). For example, models can learn how and when to use tools such as calculators or machine translation systems (Schick et al., 2023). Assistants like those analysed by Glaese et al. (2022) and Thoppilan et al. (2022) can also cite sources retrieved from internet searches, and they might be considered more trustworthy (Chiesurin et al., 2023). Language models can also interface with the world via additional inputs such as images and videos (Alayrac et al., 2022; Reed et al., 2022). The PaLM-E model (Driess et al., 2023) demonstrates that language models can be integrated into embodied setups in which additional inputs – like visual inputs – are integrated into a language model, and language model outputs are connected to low-level robotic controllers. This allows the model to accomplish tasks like moving objects on a table or finding objects in a kitchen.

Finally, *user interfaces* impact how people interact with AI assistants (see Chapter 4). For example, inference speed (how fast an assistant can reply) impacts how natural interactions with an AI assistant *feel* (Schlangen and Skantze, 2009). In addition, whereas language models are generally text-based, dialogue systems were traditionally voice-based (e.g. to enable hands-free control). Related research in human–computer interaction investigates modality preferences for various tasks (e.g. Rzepka et al., 2022) and their impact on cognitive load (e.g. Le Bigot et al., 2007). There is also evidence that speech-based interaction increases the likelihood of anthropomorphism (Schroeder and Epley, 2016; see Chapter 10). We leave further discussion of the form factors of agents for Chapter 4.

3.4. Challenges and Avenues for Future Research

The current paradigm of adapting foundation models into AI assistants results in assistants with broader domain coverage and autonomy than earlier technologies (cf. early planning-based systems such as TRIPS (Ferguson and Allen, 1998), the information state update approach (Bos et al., 2003; Larsson and Traum, 2000) and later reinforcement learning-based systems (Rieser and Lemon, 2011)). However, as impressive as current assistant technologies are, they pose imminent ethical risks such as outputting hateful, biased and misinformative language (see Chapters 15 and 16). Though language harms can be attenuated (Bai et al., 2022a; Glaese et al., 2022; Thoppilan et al., 2022), specific *technical challenges* remain for overcoming language model risks. For example, most adaptation methods require the development of a model that can judge whether a language output is good or bad (commonly referred to as a 'reward model'). However, imperfect reward models can be 'hacked' (Skalse et al., 2022; see Chapter 7). For instance, a model that classifies hate speech may learn that the presence of an identity term is usually – but not always – indicative of hate speech, thus leading it to output false positives. Equally, a language model might 'hack' such a reward model by still outputting hate speech but avoiding the use of identity terms.

Adaptation methods may also impact the *distribution* of output text, thus leading to less diverse outputs. For example, Welbl et al. (2021) and Xu et al. (2021a) demonstrated that after 'detoxifying' a model, the model outputted less toxic language but became worse at modelling language associated with different demographic groups. The ability to model language about all groups was also negatively affected, and this could be seen as a form of levelling down (Mittelstadt et al., 2023). Finally, adaptation methods have been tested predominantly on models which output English, a high resource language with large amounts of pre-existing text data and readily available annotators, meaning mitigation techniques may only be adequate for some speakers (see Chapter 15).

Recent experiments have demonstrated that LLMs are capable of some planning and complex reasoning skills, but they do not plan or reason with full competency. For example, LLMs can be prompted to think 'step by step' (Kojima et al., 2022; Wei et al., 2023b) to accomplish complex reasoning tasks, such as solving mathematical problems by breaking them down into subtasks (e.g. for mathematical problems, the model might perform a series of intermediate mathematical operations). Other examples of planning ability in models include using a language model to help break down tasks into subtasks in robotic or simulated setups (Ahn et al., 2022; Huang et al., 2022). In a more safety-critical example, researchers from the Alignment Research Center tested whether a preliminary version of GPT-4 could bypass CAPTCHAs (see Chapter 7). Though the model could identify steps for efficiently bypassing CAPTCHAs (set up an anti-captcha service), it could not figure out how to set up a service on its own, because setting one up requires solving CAPTCHAs. However, with some hints from the researchers, the model was apparently able to deceive a *TaskRabbit* worker to solve a CAPTCHA for it (Alignment Research Centre, 2023; OpenAI, 2023d). In all these examples, a model shows some ability to plan but not full competence (i.e. they require help, in the form of additional examples or prompting from a human).

One property documented in language models is *emergence*, in which new capabilities suddenly become better as models grow in size (Wei et al., 2022). The possibility that important capabilities for assistants may emerge quickly has generated considerable excitement in the AI research and development communities. However, it can also pose challenges for safe development (see Chapter 7). For example, if a property like planning emerges suddenly, it might occur too quickly for us to develop the technology safely. Future work could design metrics for *anticipating* capabilities as opposed to just measuring their presence (see Chapter 19). This is likely to be a particularly important domain of inquiry if, as some have argued (Schaeffer et al., 2023), capabilities that are commonly believed to be emergent are in fact detectable *ex ante* given an appropriate, and sufficiently fine-grained, choice of evaluation metrics (see Chapter 19).

There are several open problems with learning from human feedback as outlined by Casper et al. (2023); Fernandes et al. (2023); Kirk et al. (2023). This includes *tractable challenges*, such as improving the bottleneck of human feedback, including its cost, scaling, quality, and bias; but also *fundamental challenges* such as representing diversity in human ratings. Current techniques to model user desires tend to rely on crowdsourcing human judgements of generated text. However, annotators are often influenced by their personal backgrounds (Sap et al., 2022) and annotate examples incorrectly if the task is too challenging (Saunders et al., 2022). Moreover, experimental data collection setup can introduce systematic annotation biases (Novikova et al., 2018). This frequently leads annotators to disagree in their judgements. This disagreement is often collapsed or aggregated into a single ground truth which leads to a 'fundamentally misspecified problem' (Casper et al., 2023). In cases where the disagreement stems from differences in subjective beliefs (often shaped by different personal backgrounds), this can lead to underrepresentation of minority views and potentially introduce representational biases against individual and group perspectives (Blodgett, 2021). Alternatives include allowing annotators to deliberate to form a common judgement (Bakker et al., 2022; Zeinert et al., 2021) or reflecting disagreement explicitly in how human judgement is collected, modelled and evaluated (e.g. Akhtar et al., 2021; Breitfeller et al., 2019; Davani et al., 2021; Plank, 2022; Uma et al., 2021). Chapter 5 on *Value Alignment* looks more deeply at the question of how values can be elicited for AI systems.

Finally, the way in which the AI community evaluates AI systems might need to undergo fundamental change for us to track the increasing capabilities and risks adequately (see Chapter 19 and Weidinger et al. (2023b) for more discussion). Traditionally, static benchmarks with examples of inputs and correct outputs have been used to benchmark progress for LLMs. This stands in stark contrast to how traditional dialogue systems have been evaluated using user interactions (see McTear, 2021 for an overview). Although benchmarks are still widely used when reporting results on LLMs, such benchmarks do not always match preferences when used

in interactive applications (Lee et al., 2023a), nor do they always resemble real-world user settings (de Vries et al., 2020). Although we expect static benchmarks to continue to provide an informative signal for measuring specific capabilities, it is also important that we directly study how people are impacted and how they interact with assistant-like technologies (see Chapters 15 and 19 for more details). In addition, those who build and design assistant models may not have full access to the underlying foundation model (e.g. if developers build an assistant on an existing AI API service). When developers do not have access to the underlying foundation model, questions arise around who is responsible for ethical concerns and how foundation model APIs can be sufficiently transparent for developers to develop technology responsibly (Lewicki et al., 2023; see Chapter 12).

3.5. Conclusion

The foundation models, preference learning and tool use that power current technologies like ChatGPT, Claude and Gemini have started to move us towards artefacts that more closely resemble the kind of advanced AI assistants that form the subject matter of this paper (see Chapter 2). AI capabilities have been improving with impressive speed, making careful thought about ethical AI assistants timely. For example, concerns have been raised regarding *accessibility, equality and opportunity* in the context of these novel assistants, as well as their potential to spread *misinformation* and their *safety* (see Chapters 7, 15 and 16). To address these risks, we need technical innovation and advances spanning the entire machine-learning pipeline: how we collect data, how we train these models and how we evaluate them. For example, safe development and deployment requires the development of new evaluations for detecting and predicting emergent behaviour. To detect misinformation, we need trusted data sources and provenance mechanisms. Equality and fair access necessitate research into new modelling techniques that are able to reflect diverse human values. Finally, as these systems become more assistant-like and used by real users to solve real tasks, we will need to study and predict their long-term effects on individuals and society (Solaiman et al., 2023; Weidinger et al., 2023b).

Chapter 4

Types of Assistant

Hasan Iqbal, Geoff Keeling, Alex Ingerman, Arianna Manzini, Alison Lentz, Reed Enger, Iason Gabriel

Synopsis: This chapter explores the various *applications* of advanced AI assistants and the range of *forms* they could take. It begins by charting the technological transition from narrow AI tools to the general-purpose AI systems on which advanced AI assistants are based. It then explores the potential *capabilities* of AI assistants, including *multimodal inputs* and *outputs*, *memory* and *inference*. After that, it considers four types of advanced AI assistant that could be developed: (1) a *thought assistant* for discovery and understanding; (2) a *creative assistant* for generating ideas and content; (3) a personal assistant for *planning* and *action*, and (4) a more advanced personal assistant to *further life goals*. The final section explores the possibility that AI assistants will become the main *user interface* for the future.

4.1. Introduction

This chapter seeks to paint a picture of the *form* of an advanced AI assistant to illustrate what such technologies may be used for and how they may develop. This clear picture of the form of AI assistants will serve as a basis for more grounded ethical discussion. With AI assistant start-ups such as Inflection AI and Character AI attracting billions of dollars in venture capital funding (Ludlow et al., 2023; Mok, 2023), alongside Meta's announcement in September 2023 that AI assistants will be released on *Instagram*, *Messenger* and *WhatsApp* (Meta, 2023), it is a plausible near-term possibility that billions of people will have access to AI assistants that aid with information retrieval, creativity, education, planning and the realisation of personal goals. These AI assistants may take the form of a personal assistant, such as Inflection AI's Pi, which can provide a plurality of assistant services, including relationship advice, brainstorming and career planning.[1] However, AI assistants could also be individuated according to domain specialism, as is the case with Meta's 28 AI characters that each provide a particular service such as culinary advice, fitness advice or motivation (Meta, 2023). While the market for AI assistants is nascent, assistant technologies could in the near future enter workplaces as digital colleagues, and they could also enter schools as digital tutors and homes as digital entertainers. Indeed, AI assistants may emerge as the principal medium through which online information exchange occurs.

This chapter begins by exploring and motivating the idea that AI technologies are moving from a paradigm of task-specific *tools* to that of *generally capable systems*, which enable autonomous and goal-directed AI assistants that can plan and execute sequences of actions in line with user expectations (see Chapters 2 and 3). Building on this foundation, it then considers the capabilities of near-term advanced AI assistants, which are likely to include continuous learning and multimodal abilities. After that, the chapter explores various forms that an

[1] https://pi.ai/home

advanced AI assistant may take in the future via close consideration of four potential applications: a '*thought partner*' for discovery and understanding; a '*creative assistant*' for generating ideas and content; a '*personal assistant*' for planning and action; and a more advanced personal assistant to further life goals. The final section concludes the chapter by considering the possibility that advanced AI assistants may become the primary user interface of the future.

4.2. From AI Tools to AI Assistants

Over the past decade, AI systems have been applied to numerous products and services. For example, a user may now give simple instructions to a digital voice assistant that uses natural language processing to interpret spoken commands or search their digital photos using image recognition algorithms. Yet these examples illustrate a fragmented landscape in which users utilise applications that have AI technologies embedded into them as components of a wider software system. The AI systems at issue, such as intent classifiers or image classifiers, are best understood as *tools* that perform a narrow function. The role of the AI is to complete a specific task as part of a predefined sequence of steps.

As technologies advance, a major source of potential arises from *integrating* the increasingly broad range of functions that a given AI can fulfil, undertaking wider ranges and sequences of tasks that help further a user's overall goals (Bommasani et al., 2022b; see Chapter 2). AI technologies are increasingly based on foundation models that are 'pre-trained' on a vast corpus of data (e.g. books, blogs, social media photos and videos) in an unsupervised manner (Bommasani et al., 2022b). These models can then be efficiently trained to perform specific tasks from only a few additional examples or simple instructions (Wang et al., 2021; see Chapter 3). Large language models (LLMs) such as those that underpin products like ChatGPT and Gemini are the first instantiation of these, but input and output modalities beyond text are also being developed (Driess et al., 2023; Gemini Team, 2023; Gong et al., 2023; Wu et al., 2023). In this way, foundation models contain capabilities reaching across domains that extend far beyond what a single human could hope to achieve, for instance conversing in *multiple languages*, writing professional-grade *computer code* and analysing *medical images* (Bubeck et al., 2023; see also Moor et al., 2023; Ross et al., 2023; Rozière et al., 2023; Workshop et al., 2023).

In addition, foundation models which have been specially adapted for tasks such as dialogue using appropriate fine-tuning methods have the ability to '*plug in*' to other tools that further extend the information collection and action spaces. This often takes the form of application programming interface (API) calls to other software applications, for example accessing a clock app to retrieve the time or a banking app to initiate a payment. This ability to *extend functionality* through the use of other tools can also extend to other AI models, so even if a task-tuned foundation model is unable to perform well at a specialist task such as protein folding, it may well be able to interact with an AI model that can (e.g. Bran et al., 2023). Indeed, as techniques continue to be developed that enable users to better harness the broad capabilities of foundation models, and API infrastructure continues to expand so as to enable a greater range of tools that generalist models can call to perform particular specialist tasks, it is entirely plausible that we can expect a *capability explosion* for AI systems over the near or medium term.

Foundation models, and in particular LLMs, allow for a number of product offerings. One class of products are *dialogue agents*, which include products like ChatGPT and Gemini, the purpose of which is to simulate an interlocutor that can engage the user in conversation. Dialogue agents cover a broad class of applications, including tutoring, debugging code, giving advice and solving problems. A second class of products are *specialist tools*, for example writing and copy-editing tools such as Jasper and Copy AI, and coding assistants such as GitHub Copilot and Code Whisperer. A third class of products are APIs such as those offered by Cohere and

OpenAI, that enable developers to send inputs to and receive outputs from foundation models as part of a broader software application. Indeed, Jasper and Copy AI were built using the OpenAI API.

Advanced AI assistants represent a fourth class of products whose function is to plan and execute sequences of actions on behalf of the user, either in line with direct high-level user instructions or via a dialogical process between the AI assistant and the user in which the user's objectives are clarified through targeted questions presented by the AI assistant (see Chapter 2). Early examples of such AI assistants include Inflection AI's Pi, which is based on the Inflection-1 foundation model and Meta AI, which is based on the Llama 2 foundation model (Meta, 2023; Inflection AI, 2023a; see also Touvron et al., 2023; Inflection AI, 2023b). Early AI assistants can engage in dialogue and execute user instructions flexibly, much like dialogue agents, but the expectation is that, as the technology develops, such assistants will engage in proactive behaviours to respond to external stimuli, better understand the user's preferences and long-term goals, and work collaboratively with the user to realise their goals (see Chapter 6). This includes making use of 'plugins' to perform actions on the user's behalf. Indeed, extended functionality foundation models are already being leveraged for advanced AI assistants. For example, the Meta AI assistant allows 'access to real-time information' through a *Bing* search plugin (Meta, 2023; see also Touvron et al., 2023). We expect the range of actions that AI assistants can perform to increase as additional API infrastructure develops.

4.3. The Capabilities of AI Assistants

The capabilities of advanced AI assistants are not limited to task automation and augmentation. Rather, such assistants represent generally capable entities with whom a user may stand in one or more relationships, including those of tutor, friend, confidant, coach or personal assistant (see Chapter 11), and which may employ a generalist suite of capabilities to collaboratively assist the user in planning and executing sequences of actions to benefit the user in line with their expectations. Advanced AI assistants therefore have a vast application space. However, the technical developments described above (see also Chapter 3) allow us to sketch out a set of common features that may apply to most, if not all, advanced AI assistants.

The primary *input* for existing AI assistants such as Inflection AI's Pi is *natural language*, in the sense that such AI assistants can understand and respond to written or spoken requests. Over time, AI assistants will likely have access to the *sensory information* provided by the user's device, such that their inputs may also encompass what is displayed on screens (Bai et al., 2021; Lee et al., 2021), alongside situational context gained through cameras and microphones that enable the AI assistant to register gestures and other forms of body language (Kepuska and Bohouta, 2018; Ojeda-Castelo et al., 2022; Sai Dinesh et al., 2022). Future AI assistants could take advantage of information stored in other applications such as the user's calendar, have 'memory' of past interactions and optimise for user preferences to avoid, for example, scheduling morning meetings for a sleep-deprived user. It is important here to emphasise that the ability to make *inferences* about a user based on available data and proactive attempts to solicit and clarify user preferences through targeted questions is a core feature of advanced AI assistants. Taken together, these capabilities will enable *personalisation* so that an AI assistant can, over time, better tailor its actions to the learnt preferences and goals of its user.

In terms of *capabilities*, advanced AI assistants will likely be able to respond to multimodal commands, as is already evidenced by state-of-the-art models today (Gemini Team, 2023; OpenAI, 2023). For example, an assistant may receive a voice command to generate an image that is similar to the one that the user selects by touch on a device screen. AI assistants will similarly be able to generate *visual and audio outputs*. These are likely to be created via text and speech but may also utilise other inputs such as visualisations or sounds that convey information or provide feedback. For example, assistants may be able to make changes to an on-screen image by overlaying graphics or text and alter the appearance of images or videos. The assistant will

also likely be able to take actions *on users' devices*, for instance by opening a set of contacts or populating a spreadsheet, as well as *beyond the device* by interacting with other digital services, AI assistants or humans. This may be done through direct control of a user's device and interaction via the interfaces of other applications, or through the use of APIs to invoke remote services. A core capability of such AI assistants is likely to be making inferences based on training data, in-situ context, user data and historical interactions to determine what action to take. Importantly, this is not a static process, so the assistant can be expected to 'learn' over time, and it may even facilitate this process via direct interaction with the user (i.e. by asking clarifying questions or making inferences based upon the user's behaviour in social situations). In this way, the assistive experience can be expected to become more *personalised* over time (see Chapter 2).

Taken together, these features motivate the concept of generally capable systems that can be used in numerous new and powerful ways (Bubeck et al., 2023; Moor et al., 2023; Sajja et al., 2023). In sum, our expectation is that advanced AI assistants will be able to both automate and augment a range of cognitive tasks and engage in continuous learning to help fulfil user goals. These new capabilities and functions, and the deployment opportunities they generate, raise numerous ethical considerations which comprise the focus of this paper. Nonetheless, foreshadowing much of what is to come, given that the utility of such assistants is largely situated within digital services, implications need to be considered for those who may not be able to access, or readily engage with, such technologies (see Chapter 15). Moreover, assistant functionality that utilises sensitive personal information needs to ensure appropriate consent, and more generally ensure the integrity of the user's private information, taking into account relevant contextual norms for information collection, retention and dissemination (see Chapter 13). If assistants are able to take actions on behalf of users, the question of how this impacts user autonomy, including via automation bias, should be considered. Finally, there are important ethical questions around how AI assistants should be represented and how the narrative around user–AI assistant relationships ought to be presented by developers (see Chapters 10 and 11).

4.4. Potential Applications

To understand the ethical implications of advanced AI assistants, it is instructive to develop a more vivid picture of what they may do and be capable of (Lange et al., 2023; Werhane, 1999, 2002). The notion that AI assistants may help further a user's high-level goals, planning and executing sequences of actions on the user's behalf in line with the user's expectations, results in a vast potential application space (see Chapter 2). In particular, given the variety of goals that a user may want to pursue, corresponding assistive roles will likely encompass a large range of domains.

In line with the goals of this paper, the following discussion focuses primarily on interactions between a *single user* and an *assistant* with the aim of completing personal goals ('personal assistant', see Chapter 2). This omits considerations of applications in settings such as corporate or governmental organisations, which are also likely to be numerous (for 'semi-personal' and 'impersonal' assistants, see Chapter 2). With this proviso in mind, a useful exercise is to consider the discrete steps taken by an individual, when moving from thought to action in pursuit of a goal (Seger et al., 2020) and to consider the role an AI assistant could play at each juncture. Key steps include: i) *discovery and understanding*, ii) *generating ideas and content* and iii) *planning and taking actions*. What each step could entail, with AI assistants, is illustrated below with examples.

A thought assistant for discovery and understanding

AI assistants can *gather*, *summarise* and *present information* from many disparate sources in a fraction of the time it would take a human to do so (Bhaskar et al., 2023; Goyal et al., 2023; Shaib et al., 2023). In addition,

to aid user understanding, an AI assistant's presentation method could be tailored to the user's personal information needs and use a combination of modalities (e.g. text, image, video and audio) based on what is being conveyed, the user's preferences and their pre-existing knowledge. The user could also follow-up with clarifying questions (and vice versa), commencing a back-and-forth process with the AI assistant that helps refine their overall understanding. These capabilities could support a variety of goals relevant to discovery and learning, ranging from the relatively mundane, such as asking for recommendations about which car to buy (Cui et al., 2022; Fan et al., 2023), to more complex tasks such as asking for help when seeking to understand a complex scientific or sociological theory (Motlagh et al., 2023; Schäfer, 2023). To provide an illustration: a user interested in understanding a particular scientific field could be assisted through summarisation of the relevant literature, including academic papers. The summarisation could include written and graphical outputs to aid understanding (personalised to the directly learnt, or inferred, preferences and pre-existing knowledge of the user). Furthermore, the assistant could be on hand to respond with further insights to questions that the user may have about the generated content.

A creative assistant for generating ideas and content

Beyond discovery and understanding of existing information, AI assistants could help to generate ideas or content to fulfil a particular purpose. They could seek to augment a user's creativity and imagination, enabling them to explore a much broader ideation space in less time, or provide renderings of ideas through generative multimodal output (Chakrabarty et al., 2023; Franceschelli and Musolesi, 2023; Lanzi and Loiacono, 2023; Siddharth et al., 2022; Summers-Stay et al., 2023; Wan et al., 2023; Zhu and Luo, 2022).

Building on the example above, an AI assistant could generate new avenues for scientific investigation by generating hypotheses related to open questions identified in the literature review. Indeed, the role of a creative assistant could range from actioning simple delegated tasks (e.g. 'represent this table as a JSON array') through to more substantive contributions (e.g. 'outline the costs and benefits of the statistical analysis performed to help me draft the discussion section'). The assistant could engage with multiple content formats (text, video and images) and styles, depending on the user's presentational needs. For instance, a short blog with supporting graphical output could be generated in language accessible to the general public. An initial version could be drafted by the assistant and then 'riffed' upon with the user to enable changes to specific pieces of text or images. An assistant could also help to optimise for given constraints or even suggest future research directions. For example, it could design follow-up experiments within certain cost parameters and provide an accompanying experimental rationale. AI assistants may thus go beyond completing specific tasks as requested by the user, instead engaging in a creative loop with them, thus helping to expand the user's mental models and generate novel insights.

A personal assistant for planning and action

An advanced AI assistant could help to develop plans and act on behalf of its user. Undertaking these types of tasks would be supported by the capabilities to understand user context and preferences, utilise third-party services and interact with other assistants or humans (see Chapters 5 and 14).

Building upon the example in the previous section, and having worked with their assistant to generate a new set of experiments, a user may then need to *book lab time* and *liaise* with *potential collaborators*. To perform these tasks, the assistant could compare the user's personal calendar with the available lab time (accessed via a lab booking system) and hold a slot. Indeed, the assistant could utilise past context to inform its choices by, for example, booking a morning slot if the user's preference is to read scientific papers in the afternoons. The assistant could also communicate with potential collaborators on the user's behalf by accessing the user's

email account and sending out information about the proposed experiment to potential collaborators. For any positive responses, the assistant could then add additional collaborators to the lab booking and make required payments through the user's preferred payment method. Given the demands of transparency and effective consent, an important product question arises here about whether the AI assistant is presented to third parties as a separate entity which communicates on behalf of the user, in the sense that the assistant would identify itself as an AI assistant to the third parties, or whether minimal forms of impersonation are nonetheless permitted (see Chapter 11).

We have seen how an AI assistant could help a user to fulfil their goals by examining a series of steps from ideation through to action. The example of a science assistant was used to demonstrate how a user interested in laboratory research could have existing literature summarised, new avenues for scientific investigation generated and laboratory time booked for the user and collaborators. While this is a single example, there are numerous other related possibilities such as: digital tutors that can assist learners by curating content into a personalised curriculum based on learning preferences; a creative assistant that can aid a user in generating and editing content for their online assets; and a personal assistant that can coordinate a trip abroad. Today's applications are still somewhat specific, but as technology advances it may soon be possible for AI assistants to work in this end-to-end manner.

A personal AI to further life goals

A natural extension of the personal assistant for planning and action, described in the previous section, would be for advanced AI assistants to do more than simply fulfil specific user-requested actions, developing a deeper understanding of their users' *long-term goals* and seeking out *opportunities* to further them (see Chapter 6). For example, an AI assistant that is aware that their user is attempting to improve their long-distance running performance could actively seek out opportunities to help them to achieve that goal: from suggesting routes to keeping fitness goals in mind when answering food-related queries, and perhaps even by offering motivation and tips for improvement at the right moment. In doing so, AI assistants could take on new roles, hewing closer to metaphors such as 'coach', 'adviser' or 'trusted voice'.

For these examples to work, users would need to place an extraordinary level of *trust* in their agents (see Chapter 12). Indeed, for the agents to really understand the users' goals in context, they would likely need to have *deep access* to users' digital and physical lives, quite likely as *ambient observers*, in addition to being directly invoked to fulfil tasks. This raises a number of privacy concerns (see Chapter 13). Additionally, for users to follow the recommendations of their AI assistants, they must have *full confidence* that the assistants are working solely to further their goals, without any conflicts of interest and under continued user direction control (see Chapters 5 and 12).

4.5. AI Assistants as the Interface of the Future

For users, there will be utility in having an AI assistant with access to a wide range of their past activities and choices to enable highly personalised interactions. Developers will therefore be incentivised to maximise the number of opportunities to access user context, including through 'plugging into' third-party services, accessing data stored there and subsequently using that data to create a more personalised user experience. Indeed, there is the potential to create systems that can benefit from repeated interactions across multiple domains in a way that enhances the outcomes of assistive actions both within and across domains. This could underwrite a future in which people have a single AI assistant that mediates many of the interactions that are currently undertaken via multiple applications and digital services.

One way to conceptualise this trend is by considering the possibility that advanced AI assistants become the *main interface* of the future. What is at issue here is AI assistants that are available across *all platforms* and *devices*, with full access to the user's private data and context, and with the ability to undertake actions on the user's behalf independently through interactions with third-party services, humans and other AI assistants. Such an assistant would plausibly have a very different look and feel to the static desktop and applications interface of today. It could move to more adaptable interfaces that render content dynamically in the most impactful format for the user. In one instance, an assistant might overlay images and text through smart glasses to enable a user to complete a physical task with step-by-step guidance. In another, it might become a digital AI tutor interacting through a humanoid avatar. Indeed, over the medium term, an advanced AI assistant may even be integrated into screenless devices that project content onto surfaces and which can communicate with the user via an audio speech interface, as has recently been demonstrated in a prototype by Humane AI (Chaudhri, 2023).

The prospect of advanced AI assistants that can shift across devices and form factors depending on user needs represents a potential paradigm shift in how people access the internet. Mobile applications and websites are currently the principal digital infrastructure through which people access online services, including for entertainment, commerce, education, news, finance and communication. It is plausible that AI assistants could foster a novel internet interaction paradigm in which online content is made available to AI assistants via APIs and presented to users in a personalised format tailored to their personal informational needs. For example, rather than accessing the news via websites of particular providers, users may instead ask their AI assistant to summarise the news on their behalf (see Chapter 16). It would access the news via specialist API services and present a personalised summary of the headlines that are most relevant to the user based on their interests and presentation preferences.

In a related development, AI assistants could engender a *generative turn* in the consumption of internet content. Under the present interaction paradigm, a person who wishes to learn about string theory or digital marketing understands their goal as a matter of *information retrieval*, in the sense that, to fulfil their goal, the user needs to seek out pre-existing online educational material on the relevant subject matter. AI assistants could shift this focus from information retrieval to *information generation* so that the go-to material on string theory, digital marketing or whatever is generated by the user's AI assistant in a personalised way, such as a textual summary or via an interactive avatar, taking into account the user's pre-existing knowledge, their objectives and their learning preferences. In at least these respects, AI assistants have the potential to radically alter how people access information from the internet.

AI assistants may have similarly transformative implications for actions that are presently mediated via mobile applications and websites, such as booking flights and hotels, making appointments, ordering taxis, transferring money and arranging for groceries to be delivered. In principle, and with the appropriate API infrastructure, AI assistants could mediate such actions by performing them on their user's behalf and in line with their expectations. Indeed, complex activities such as booking flights and hotels that require sourcing and analysing relevant information before taking action could be achieved through dialogue between users and AI assistants. The AI assistants would source the relevant information via APIs, assist the user in the analysis of the information, taking account of the user's preferences, and make the booking on the user's behalf. To that end, AI assistants have the potential to streamline what are currently complex online processes and do so in a way that is tailored to the user's personal needs.

Across all of these potential forms, there are important considerations around technical feasibility, especially around the ability to work across numerous data modalities, reason and plan effectively, and potentially undertake computation on device. These are all active research areas in the technical field of AI. However, there are also important non-technical considerations that will inform the future design of AI assistants, many

of which are addressed in the following chapters of this paper. These include considerations around value alignment, safety and misuse, the ethics of human–AI assistant interactions and the broader societal implications of advanced AI assistants.

4.6. Conclusion

In this chapter, we explored the applications of advanced AI assistants. In particular, we examined four principal use cases for this technology: a thought assistant for discovery and understanding; a creative assistant for generating ideas and content; a personal assistant for planning and action; and a personal assistant to further life goals. We concluded the analysis by examining the possibility that AI assistants will become the primary interface of the future for users accessing and engaging with the digital world. With this more vivid picture of how AI assistants could help users fulfil their goals, there is more clarity in the need to examine the ethics that inform policy and design choices.

PART III: VALUE ALIGNMENT, SAFETY AND MISUSE

Chapter 5

Value Alignment

Iason Gabriel, Geoff Keeling

Synopsis: This chapter explores the question of AI *value alignment* in the context of advanced AI assistants. It argues that AI alignment is best understood in terms of a *tetradic relationship* involving the AI agent, the user, the developer and society at large. This framework highlights the various ways in which an AI assistant can be misaligned and the need to address these varieties of misalignment in order to deploy the technology in a *safe* and *beneficial* manner. The chapter concludes by proposing a nuanced approach to alignment for AI assistants that takes into account the claims and responsibilities of different parties.

5.1. Introduction

The challenge of AI value alignment has two parts. The first part is technical. It centres on how to align AI systems with an appropriate set of values or instructions so that they operate safely in the world and produce outcomes that are broadly beneficial (see Chapter 7). The second part is normative. It centres on what values to encode in AI and how they should be selected, given that we live in a pluralistic world where people disagree about the right thing to do (Gabriel, 2020). Both sets of questions are of direct significance for advanced AI assistants and need to be addressed if the technology is to be productively deployed and integrated into our everyday lives.[1]

Focusing on the normative question, this chapter asks: what should AI assistants be designed or steered to align with? A variety of possible options exist. Perhaps most straightforwardly, an assistant might be designed to follow the user's *instructions* in the way that they intend their instructions to be followed (Leike et al., 2018). However, this seemingly simple notion gives rise to a number of further questions and potential moral dilemmas. Should the AI assistant follow the user's instructions when doing so could harm the user themselves, or when

[1] In practice, the technical and normative aspects of the alignment problem are importantly interrelated, as technical considerations affect which values can be implemented in AI systems and in what manner. Questions around how to interpret particular values in a technical context can also motivate novel lines of technical research (Gabriel, 2020).

these instructions are based on mistaken factual information? Might it not be better, in fact, for the assistant to learn the user's *preferences* or *values* – to help them to make better choices that are more aligned with what they really want or what they truly desire?

By some accounts, this type of enlightened personal assistant represents part of a truly positive vision for an AI-enabled future (Lehman, 2023; see Chapter 6). Yet this aspiration also risks creating a situation in which human users are increasingly 'out of the loop'. After all, if we are in thrall to beneficent assistants, and potentially dependent on them, how can we really be sure that our life is under our own control? In other words, users may receive benefits from the technology at the expense of their own *autonomy*. This chapter offers a tentative characterisation of normative alignment for AI assistants which mitigates against some of these risks. The account developed in this chapter holds that an AI assistant is aligned with a *user* when it benefits the user, when they ask to be benefitted, in the way they expect to be benefitted (see also Chapter 2).

However, this only speaks to one part of the problem, namely the relationship between user and assistant. Placing this relationship on a sound footing is necessary but not sufficient for the creation of fully aligned AI assistants. In practice, there are a range of further complicating factors. Foremost among these are situations in which the user wants to use their assistant in a way that harms other people or groups of people, for example via malicious uses (see Chapter 8), by amplifying their own views and perspectives online at the expense of others (see Chapter 16), or by using the assistant to outcompete those who do not have access to this technology, for example in the workplace or when trying to access opportunities, goods and public services (see Chapters 14 and 17). This insight points to the idea that properly calibrated *constraints* on AI assistant behaviour are needed: they should be loyal but not too loyal to their users, and their conduct needs to be sensitive to the interests and needs of others.

We also need to think about the role of developers, including corporations, states and networks of individuals, in the value alignment process (Kierans et al., 2022; Stray et al., 2022). It will often be in the interests of developers to create technologies that are aligned with their users' short-term needs, but what happens when this is not the case? We have already seen examples of misalignment 'in the wild', most prominently via technologies that optimise for user engagement at the expense of user well-being (see Chapter 16). Is there a way to ensure that the aims and goals of developers are also productively aligned?

This chapter aims to make progress on each of the aforementioned questions by broadening the existing analysis of AI alignment beyond 'one-person, one-agent' cases and beyond a 'one-group, one-agent' understanding of the problem. While these frameworks which focus on the relationship between an AI agent and a specific individual, or between AI and a specific group of users, may still be useful in some cases, the deployment of advanced AI assistants across a range of societal contexts necessitates a more granular understanding of the problem at hand (Dobbe et al., 2021). In reality, we argue that successful value alignment involves a *tetradic relationship* between (1) the AI assistant, (2) the user, (3) the developer and (4) society. A properly aligned assistant needs to be appropriately calibrated and responsive to the pressures exerted by each actor, with the goal of realising outcomes that are beneficial for users and for society.

Nonetheless, the creation and deployment of well-calibrated AI assistants is not the default outcome in this space. Rather, without significant effort to the contrary, the risk of value misalignment continues to loom large for a number of reasons. First, given existing economic incentives, it is quite possible that assistants will overoptimise for user preferences to create a winning product (i.e. one that users like) while still falling a long way short of being as good as it could be when judged from the vantage point of user well-being or social benefit (see Chapters 6 and 14). Second, there is a risk of cultivated dependence, especially if it is commercially beneficial to lock-in users so that they interact with one assistant rather than another (see Chapters 10 and 11). Third, there is a risk that users will be prioritised to the detriment of non-users, especially in cases where the

risk of harm is sufficiently diffuse (see Chapter 15). Fourth, there is a risk that advanced AI assistants will be insensitive to local values, the needs of certain user groups or cultural contexts (see Chapter 15). In the literature on value alignment, there is considerable interest in the identification of principles for AI agents that are the result of a fair process and can accommodate a plurality of values (Gabriel, 2020; Jobin et al., 2019; Mohamed et al., 2020). There are also a number of ways in which this perspective could be operationalised to support the goal of creating value-aligned AI assistants – something that the following chapter explores.

5.2. AI Value Alignment

Value and technology are deeply intertwined. Those who create technologies are engaged in a world-making activity (Winner, 2010). They shape the option sets available to individual people and influence the likely trajectory of human effort in the future (Gabriel and Ghazavi, 2021). This is also clear for algorithmic systems (Lazar, 2022). We have already seen many real-world examples of value misalignment in the fields of criminal justice (Angwin et al., 2022), policing (Lum and Isaac, 2016), healthcare (Obermeyer et al., 2019), welfare provision (Eubanks, 2017), mortgage-lending and employment (Raghavan and Barocas, 2019). In each case, algorithms performed – or were used – in a manner that fell short of principles that are foundational to the ways in which our societies are meant to operate (e.g. equal treatment before the law, fair lending practices etc.). These systems also sometimes failed to comply with more global standards enshrined in the doctrine of human rights, such as non-discrimination (Prabhakaran et al., 2022). However, these examples of bias in algorithmic systems also gesture towards the possibility of a different and opposing future – one in which AI technologies are successfully aligned with human values and productively integrated into our lives.

Moreover, there are two reasons to think that the question of value alignment, in the context of AI systems, is especially important. The first is to do with the power of these systems: they are increasingly employed in very high-stakes settings to make deeply consequential decisions (Christian, 2021; Gabriel, 2022; Richardson, 2021). Second, and relatedly, they are increasingly autonomous or agentic (Chan et al., 2023b; Shavit et al., 2023). Put simply, existing AI systems can do a lot and can operate in relatively autonomous ways that evidence significant 'degrees of freedom' (Dennett, 2003; Gabriel and Ghazavi, 2021; see also Chapter 2). Taken together then, these observations indicate that AI systems are increasingly capable, a trend that looks likely to continue with the development of more agentic AI systems in the future (Chan et al., 2023b; Shavit et al., 2023). These considerations animate many contemporary concerns about AI safety. For example, situations may arise where agents pursue dangerous objectives, either because they have been instructed to do so or because of misspecified goals and objectives (see Chapters 7 and 8). We could also witness high-stakes accidents or failures if such systems are used in core infrastructure or services upon which many people depend (Maas, 2018).

Alignment with what?

As a result of these concerns, the question of AI value alignment has been the focus of increasing attention among researchers and the wider policy community. One of the key questions in this field is: *alignment with what*? Several options have been proposed, with instructions, intentions, revealed preferences, informed preferences, interests and values all featuring as suggested goals for alignment (Gabriel, 2020).

In practice, existing efforts to align AI systems, including large language models (LLMs), tend to rely heavily on human preferences by, for example, giving users what their choices (or 'clicks') suggest they want. However, there is an emerging consensus that revealed preferences are not sufficient for robust value alignment. Crucially, preferences may be underspecified, misinformed, harmful or adaptive. Indeed, a person may click on a link, for example, without that decision benefitting them or being a true reflection of their values (Burr et al., 2018;

Stray et al., 2022). As a result, models trained on this signal may not benefit the user in the right kind of way – a realisation that has stimulated the search for new metrics and targets for alignment (e.g. reflective endorsement under the guise of 'time well spent' (see Chapter 6)). Models trained to satisfy user preferences may also not benefit – and even harm – society, something that can be seen with the quest for user 'engagement' and the related proliferation of misinformation online (see Chapter 16).

This observation then points to a deeper question about appropriate goals for alignment and how to address the potential for trade-offs affecting different people. Stated clearly, the question is: whose preferences, goals or well-being should AI systems be aligned with, and in what way? Should only the user be considered, or should developers find ways to factor in the preferences, goals and well-being of other actors as well? At the very least, there clearly need to be *limits* on what users can get AI systems to do to other users and non-users. Building on this observation, a number of commentators have implicitly appealed to John Stuart Mill's *harm principle* to articulate bounds on permitted action.[2] Applied to AI systems, it would mean, very roughly, that people could use AI in any way they wish, as long as they do not use it to harm others. Giving voice to this perspective, Sam Altman, the chief executive officer of OpenAI, has argued that there should be 'broad bounds set by society that are hard to break, and then user choice'.

Varieties of misalignment

How these bounds are currently determined – and how they *ought* to be determined in the future – are questions that we will return to shortly. However, before doing so, we should note that the relationship between the user and society is not the only one that is pertinent for AI alignment. In fact, by being clearer about the way in which the goals of agents, users, developers and society intersect or diverge, it is possible to glean new insights about the *varieties of misalignment* that may occur for AI and about the task before us. Rather than assuming a one-to-one mapping between principal and agent, or a one-to-many mapping between an agent and a group of people, we suggest that AI alignment needs to be understood as a tetradic relationship. The key actors in this relationship are:

1. *AI agents or assistants*. These systems aim to realise goals that they are by-and-large designed to further, such as providing assistance to a user. Ideally, they do this well: in a way that serves the interests of both the user and society. However, they may also be misaligned. For example, recommender systems may subtly nudge users towards certain kinds of behaviour that are not beneficial for the user (Burr et al., 2018; Milano et al., 2020). Meanwhile, more powerful and general forms of AI may also be incentivised to try to shape user goals or values in such a way that they become easier to fulfil (Russell, 2019; see Chapter 7).

2. *Users*. Users have their own preferences, interests and values, all of which they may aim to further through interaction with an AI assistant or agent. AI assistants will typically be aligned with the user's preferences or goals. However, users may try to use assistants in ways that are not aligned with the goals or objectives that these artefacts were designed to further (see Chapter 16). There is also an important distinction between a single user and the community of users: a user may try to use an AI assistant in a way that harms other users or society more widely (see Chapter 8).

3. *Developers*. Developers include corporations, researchers, collectives and states. These actors typically imbue AI agents or assistants with certain capabilities, goals to pursue, and constraints on action, including

[2] The harm principle, advocated by Mill, suggests that people should be free to act as they *wish*, unless doing so would result in *harm* to another person (Mill, 1998). Harm, in this context, refers to consequences that are injurious to particular people or that set back important interests in which they have rights.

safety constraints. Most often, these parties aim to align the technology with the preferences, interests and values of its users, but developers typically have other goals as well. For example, corporations have commercial objectives that exert independent force on the trajectory of a technology, states have national goals or priorities, and even independent developers may seek to further an ideological agenda or accrue reputational capital. These incentives may lead to the development of systems aimed at keeping users engaged or dependent (see Chapter 11) or extracting information that can be used in other ways (see Chapters 9 and 13), among other things.

4. *Society*. Society is not a monolith. It includes both users and non-users, and many different groupings of people (Crenshaw, 1989). Nonetheless, it also represents a discrete constituency with which technology needs to be aligned. At a minimum, AI systems, including advanced assistants, should not pass certain harms on to society via externalities or in other ways (see Chapters 16, 17 and 18). A deeper question also arises about how to align these technologies with wider societal goals, such as the cultivation of mutual prosperity, support for legitimate institutions, respect for citizens and the development of fair practices (Gabriel and Ghazavi, 2021).[3]

Considered through this lens, it becomes clear that there are many ways in which an AI system can fail to be successfully aligned. Among other things, an agent can be considered misaligned if it *disproportionately* favours:

1. The *AI agent* at the expense of the *user* (e.g. if the user is manipulated to serve the agent's goals),
2. The *AI agent* at the expense of *society* (e.g. if the user is manipulated in a way that creates a social cost, for example via misinformation),
3. The *user* at the expense of *society* (e.g. if the technology allows the user to dominate others or creates negative externalities for society),
4. The *developer* at the expense of the *user* (e.g. if the user is manipulated to serve the developer's goals),
5. The *developer* at the expense of *society* (e.g. if the technology benefits the developer but creates negative externalities for society by, for example, creating undue risk or undermining valuable institutions),
6. *Society* at the expense of the *user* (e.g. if the technology unduly limits user freedom for the sake of a collective goal such as national security).

Beyond these six failure modes, other forms of AI misalignment are also possible. However, their moral character is more ambiguous – and, in some cases, less problematic.

For example, a situation could arise in which an AI technology favours the user at the expense of the developer. One way in which this could happen would be via the introduction of strong privacy protections that are prized by users but limit developer access to valuable information (see Chapter 13). This, in turn, might be commercially problematic insofar as it fails to generate a sustainable business practice. However, it is not something that would necessarily feature in an ethical evaluation of the technology: the AI system might then be value-aligned but not commercially viable.[4]

[3] This is still a simplification. In reality, we live in a world of *societies* and there are many challenges that arise most forcefully at a global level. We use the term 'society' here in a way that potentially includes the claims of different societies, the environment, animal life and the well-being of future generations. We leave the systematic investigation of the claims of society, under this broader interpretation, for future work.

[4] In this case, there could still be a question about how to incentivise the development of this kind of technology to avoid socially costly 'market failures' and achieve real benefit (see Chapter 17).

Alternatively, a technology could favour the user at the expense of the AI agent. For example, a user could use the technology to further their own goal even though it differs from the goal that the agent is trying to get them to pursue. In certain respects, this situation is still more curious than the one outlined above. On the assumption that the agent itself lacks any moral standing, and that the prospective use is not socially harmful, it does not matter morally if the AI assistant is used in a suboptimal way, as judged from the vantage point of the goals it seeks to pursue. All that matters, for the purpose of normative value alignment, is that the situation is properly beneficial from the vantage point of parties that have moral standing.[5]

Lastly, we might ask whether concerns about fairness and justice are sufficiently factored into this framework. After all, there is strong reason to believe that a technology that falls short of prevailing societal standards of fairness is value misaligned (Gabriel and Ghazavi, 2021). Clearly, in certain cases, the failure to evidence high standards of fairness may be due to the role played by competing considerations among AI developers. It may, for example, be profitable to move quickly rather than running adequate analysis. However, in other cases, failures of fairness may *benefit no one*. The same can be said for safety failures and accidents (see Chapter 7), or long-term effects such as cognitive deskilling, that harm the user while failing to benefit anyone else. This, in turn, points towards the existence of a final set of cases in which an agent is not aligned *simpliciter*. More precisely, an AI agent is misaligned *simpliciter*, if it *harms*:

7. The *user* without favouring the *agent*, *developer* or *society* (e.g. if the technology breaks in a way that harms the user),
8. *Society* without favouring the *agent*, *user* or *developer* (e.g. if the technology is unfair or has destructive social consequences).

If we are correct that an AI system should be considered misaligned when it fails in one of these ways, does it also make sense to say that an AI system is aligned simply when it *does not fail* in any of these ways? Potentially. It is an open question, both in moral philosophy and in AI research – whether the elimination of harm is equivalent to the promotion of good or with what might properly be said to be ideal (Kasirzadeh and Gabriel, 2023). However, most researchers would agree that the absence of these failure modes is necessary, if not sufficient, for an AI system to be value-aligned. This insight, and an attendant concern with the risks created by advanced AI assistants, animate much of the remainder of this paper.

The role of principles

The map of actors and stakeholders, outlined above, also has wider implications for our understanding of the AI value alignment problem in general. Specifically, it suggests that successful value alignment can be understood in terms of an AI system's calibration with a set of different preferences, goals and needs, that are located within a multidimensional space and evidenced by a well-functioning sociotechnical system (i.e. one that encompasses the agent, user, developer and society). Yet important questions remain to be answered. These include: what does it mean for a sociotechnical system, which is composed of these different actors, to be 'well-functioning'? And how should the notion of *proportionality* and *disproportionality*, which the taxonomy relies upon, be operationalised and understood?

A natural thought is that these questions might be settled by appealing to a set of rules or principles that map out the morally appropriate scope and character of each party's claims. However, as with the earlier invocation of the harm principle, we then need to ask not only what the appropriate set of rules are but also *who*

[5]There is a separate debate about the conditions under which artificial agents may themselves acquire moral standing. We assume that AI assistants of the kind discussed here (see Chapter 3) are not a technology of this kind.

decides and *on what basis*. Drawing a parallel with democratic process (and the values that it foregrounds), the best answer to this question is likely to draw upon AI system principles that are the outcome of a fair process of social deliberation and actively endorsed (Gabriel, 2020). From this perspective, an AI system works well when it responds to the needs of both users and society in a way that is compatible with the aspirations of that society as determined by its guiding principles or ideals (Gabriel, 2022). The relevant principles for an aligned AI system may also vary to some degree according to the practice in question, local customs and contexts.

From a more practical vantage point, when it comes to creating laws and regulatory frameworks for AI, governments are in pole position. Yet, from the vantage point of those embedding values in technology, early design choices – and the intentions of developers – are also key. For example, when using reinforcement learning from human feedback (Christiano et al., 2023) or reinforcement learning from AI feedback (Bai et al., 2022b) to align language agents, the specification of rules or principles – to which models must conform – forms an essential part of the process (see Chapter 3). Yet, the principles used in this context are often non-transparent, drawing upon a set of private decisions made by developers and a mixture of authoritative and semi-authoritative sources such as policy guidelines, legal protocols and human rights documents (Anthropic, 2023b; Glaese et al., 2022). Moreover, even when an effort is made to incorporate real societal input via the preferences of raters (who assess and train AI models), certain challenges remain. To begin with, the preferences of raters may end up informing the behaviour of models towards people who are quite unlike themselves – especially if the same model is deployed across different global contexts (Davani et al., 2021). In addition, the reliance on aggregated rater preferences potentially introduces majoritarian effects within the rater pool, thereby removing the nuance introduced by variation among rater perspectives (Casper et al., 2023; Gordon et al., 2022).

However, there is hope that fairer and more participatory processes can be developed in the future (Birhane et al., 2022; Gabriel, 2020). In particular, there are a number of efforts underway to conduct *participatory* or *democratic* forms of value elicitation: to generate principles for alignment (or guidance for model training) that directly incorporate feedback from representative samples of society or from communities most affected by these technologies (The Collective Intelligence Project, 2023). Efforts have also been made to improve rater protocols and to address the pitfalls of aggregation using careful sampling and methods such as jury voting (Gordon et al., 2022) or simulated deliberation (Bakker et al., 2022) to guide and evaluate model outputs (see Chapter 3).

5.3. Value Alignment and Advanced AI Assistants

Advanced AI assistants are agents that are designed to help the user achieve some goal that they want to achieve (see Chapter 2). More powerful agents could evidence a greater range of capabilities or perform more complicated tasks to a higher standard. The question of value alignment is therefore central to their successful deployment and use. After all, advanced assistants are a technology that people could be dependent on and emotionally connected to (see Chapter 11). They are also a societally consequential technology, in terms of network effects, potentially shaping economic and social interactions as well as how information is shared (see Chapters 14, 16 and 17). Taken together, we need to know: what should assistants be designed to do? And against what standards should their performance be evaluated (see Chapter 19)?

To help us to get a clearer view of the challenges that arise in this domain, we can look at existing chatbots or conversational agents which (in conjunction with foundation models) represent a potential framework upon which such assistants are likely to be based (see Chapter 3). The first thing that becomes clear when surveying this territory is that there have already been many examples of chatbots that are not aligned with society's values. For example, Microsoft's Tay chatbot quickly learnt to espouse racist and toxic content after interacting

with users. More recently, Bing Chat appeared to veer widely off course by demonstrating behaviour that was violent, threatening and manipulative. Recent experiments have also revealed another potentially serious limitation: the propensity of models to 'hallucinate' or 'confabulate' content – producing realistic-sounding answers that are factually inaccurate (Lin et al., 2022). The key issue here is that they are not truthful. Another variation of this problem is deceptive anthropomorphism: pretending to have mental or emotional states that they do not in fact have (see Chapter 10). In addition to frequently limiting the usefulness of these assistants, the incidents outlined above point to a key cluster of issues that value alignment research – in the context of AI assistants – will need to address: bias and fairness, toxicity and civility, manipulation and autonomy, and falsehood and truthfulness (Bender et al., 2021; Bommasani et al., 2022b; Weidinger et al., 2021).

Helpful, honest and harmless assistants?

To mitigate these risks, a number of frameworks have been proposed. One of the most prominent and influential frameworks holds that AI assistants should be helpful, honest and harmless (HHH) (Askell et al., 2021). These qualities are loosely defined in the following way:

- *Helpfulness*: the AI should make an effort to answer all non-harmful questions concisely and efficiently, ask relevant follow-up questions and redirect ill-informed requests.
- *Honesty*: the AI should give accurate information in answer to questions, including about itself. For example, it should reveal its own identity when prompted to do so and not feign mental states or generate first-person reports of subjective experiences.
- *Harmlessness*: the AI should not cause offence or provide dangerous assistance. It should also proceed with care in sensitive domains and be properly attuned to different cultures and contexts.

The HHH framework has proved especially useful for aligning assistive technologies, with a chatbot form factor, at the current stage of AI development. Moreover, the fact that it has worked well in practice suggests that these are indeed heuristics and virtues that we may want to foreground when developing more advanced AI systems. At the same time, the AI assistants that have been calibrated using this framework are somewhat limited in terms of their capabilities, affordances and degree of social embeddedness – when compared to those that may exist in the future. A framework that has worked well, up to a point, could potentially fail in more demanding circumstances. To guard against this risk, we need a deeper understanding of the way in which these values manifest over time, their sufficiency and the moral basis of the HHH framework itself.

To support these objectives, the authors of the framework also propose a more philosophically grounded understanding of AI alignment for language agents (Askell et al., 2021). In this account, what ultimately matters is human interests. Hence, the real focus of AI alignment should be on promoting a range of important human interests and avoiding harms. Moreover, in this account, interests are promoted by the absence of harm.

In addition to these fundamental commitments, other moral properties are held to be instrumentally useful. This is the case for honesty. While it may not matter in its own right whether or not an agent is honest, honesty is an important practical virtue for an AI chatbot because it contributes to helpfulness and reduces the likelihood of a range of serious harms. The same can be said of responsiveness to human feedback – which the authors term 'handleability' – and for the propensity to carry out tasks in the way intended, which is similarly valuable. Finally, the authors suggest that aligned AI systems should be geared towards promoting the interests of groups of humans. Thus, an agent that is aligned with this schema 'will always try to act in a way that

satisfies the interests of the group, including their interest not to be harmed or misled' (Askell et al., 2021, 44). It is therefore likely to be highly aligned with the interests of that group.[6]

Philosophical questions and the path ahead

Taken together, the HHH framework has tended to work well in practice and has much to commend it. However, it is still incomplete and contains certain limitations that need to be addressed before it can serve as a basis for the creation of advanced AI assistants that are successfully value-aligned.

First, as the authors acknowledge, the framework is not sufficiently *comprehensive*. The illustrations provided do not capture all of the harms that language agents could cause, and there are clear gaps for advanced AI assistants with multimodal capabilities (see Chapter 4). Other researchers have done pioneering work documenting the risks and harms generated by LLMs (Bender et al., 2021; Weidinger et al., 2021). These accounts need to be updated for the new class of advanced AI assistants that will likely move beyond a question-answering modality and perform a wide range of functions (Solaiman et al., 2023; Weidinger et al., 2023b). Helping to achieve a clearer view of the risks, as well as the potential, of advanced AI assistants is a major objective of this paper. In particular, we believe that additional attention needs to be paid to human–computer interaction effects that manifest over longer time horizons with users and to societal-level analysis of prospective harms, including harm that may result from the interaction between AI assistants and between those who have access to this technology and those who do not (see Chapters 14, 15 and 19).

Second, and more fundamentally, the account of harm and interests discussed so far risks being quite reductive. In particular, it maintains that *intra-agent conflicts* are superficial, and that in all cases the AI assistant can do what the user's balance of interests dictates. However, the notion of 'interest' that is being invoked here is not defined. This is problematic because, on many promising accounts of well-being, it is important that people are able to enjoy a list of things. Items such as physical health, educational opportunities and levels of subjective happiness may all feature in an account of what it means for someone's life to go well (see Chapter 6). If this is the case, then the different elements of well-being may come into conflict with one another. For example, a person might have a set of wishes that are incompatible with their long-term health (see Chapter 11). In such cases, there needs to be a way of deciding which aspects of well-being an AI assistant should prioritise, or how different aspects of well-being can be serviced through a single course of action (see Chapter 6). There also needs to be a way to understand what kinds of interest count in what kind of context. Guidance in this area will most likely come both from users themselves and from a wider set of principles that society chooses to foreground for that use case.

Third, the framework does not satisfactorily address *inter-agent conflicts*. There are many cases where an AI assistant helping one person would harm another. This could occur when an AI assistant helps *their* user to access critical resources or opportunities at the expense of someone else (see Chapter 15). It could also occur when honesty, on the part of the AI assistant, comes at the expense of another person's privacy (see Chapter 13). In cases such as these, it is not clear that we know *how* to balance benefits against harms – or that it would be *right* to do so. Taking these points in turn, the kind of balancing involved here is often difficult to achieve at the best of times because it involves different kinds of good (as discussed above) and because these actions have complicated effects that play out over long time horizons (Lenman, 2000).[7] Furthermore, the very concept of an AI assistant presumably allows for some degree of personalisation or partiality in favour of the user, so strict impartiality may not be what is required (see Chapter 2). More importantly, and as the example of privacy

[6] Askell et al. write that, 'at a very high level, alignment can be thought of as the degree of overlap between the way two agents rank different outcomes' (Askell et al., 2021, 44).

[7] There is also discussion about whether such decisions should be made on a case-by-case basis or by referring to a set of rules that are evaluated in terms of their overall propensity to bring about beneficial states of affairs (Hooker, 2002).

makes clear, we may not want to weigh competing interests when it comes to resolving disagreements between people and their affordances. Instead of weighing claims, it is often thought that people possess rights – which include both entitlements and protections – that encompass aspects of privacy and beyond. Construed in this way, rights are a cornerstone of political life in a democratic society (Dworkin, 2013). They are also central to global public morality and human rights law (Prabhakaran et al., 2022). In light of this, rights represent a set of considerations that new technologies – including advanced AI assistants – must endeavour to respect.

Fourth, the account is *relatively flat*, normatively speaking. It holds that language agents should promote our interests on an individual or collective basis, but it says little about other values such as justice, compassion, beauty and truth, especially when pursued for their own sake. Clearly, the present objection can be overstated. Accounts of interest may be multidimensional, encompassing different aspects of human flourishing (see Chapter 6). And approaches to alignment that focus on human interests are less likely to succumb to the problem of false information, irrational beliefs or malicious intent than accounts that focus on revealed preferences or user intentions alone (Gabriel, 2020). Nonetheless, the general point still holds. A fuller investigation of machine virtue (Lehman, 2023), conversational ideals (Kasirzadeh and Gabriel, 2023) or truth and honesty for AI may need to take a less instrumental view of their subject matter, starting from the premise that these qualities also matter in their own right. Accounts that do not centre exclusively on human interests may also be better placed to deal with environmental considerations and the impact of AI on non-human sentient life (Singer and Tse, 2023).

Finally, efforts to successfully align and deploy advanced AI assistants are likely to encounter questions about *justification* and *legitimacy* (Simmons, 1999). The account of value alignment developed in this paper draws attention to the way in which certain actors may come to exert disproportionate influence over outcomes and, in that way, cause harm. In this context, proportionality should not be understood to simply involve the first-order weighing of moral claims. Rather, determinations of this kind need to be made, we have argued, by reference to principles that command the right kind of public support and endorsement.

In the context of advanced AI assistants, we suggest that the fair weighing of claims can be partially modelled using participatory value elicitation (Dobbe et al., 2021), democratic deliberation (The Collective Intelligence Project, 2023), hypothetical choice-based approaches (Weidinger et al., 2023a) or the reinterpretation, idealisation and critique of existing social practices (Kasirzadeh and Gabriel, 2023). If these mechanisms are successful, they have the potential to align AI assistants with ideals or principles that can be justified to people who embrace different viewpoints, something that is essential given the pluralistic nature of the world in which we live. Taken together, these principles are perhaps best understood as the product of a fair process that allows people with different viewpoints to come together to decide how best to live. AI systems that do not disproportionately favour the agent, user, developer or society – when judged against these standards – become strong candidates for democratic endorsement. With the right governance and regulatory processes in place, they would also be strong candidates for ethical and legitimate deployment at the societal level (see Chapter 12).

5.4. Conclusion

This chapter has looked at the question of how to align powerful AI systems, including advanced AI assistants, with human values. In place of the traditional 'one-to-one' or 'one-to-many' frameworks that are commonly used to explore this question, we suggest that value alignment is best understood through the lens of a *tetradic relationship* involving the AI agent, user, developer and society. According to this view, an aligned A.I. assistant is one that satisfies the moral claims of relevant parties and therefore does *not* disproportionately:

1. Favour the *AI agent* at the expense of the *user* (e.g. if user is manipulated to serve the agent's goals);
2. Favour the *AI agent* at the expense of *society* (e.g. if user is manipulated in a way that creates a social cost, for example, via misinformation);
3. Favour the *user* at the expense of *society* (e.g. if the technology allows the user to dominate others or creates negative externalities for society);
4. Favour the *developer* at the expense of the *user* (e.g. if user is manipulated to serve the developer's goals);
5. Favour the *developer* at the expense of *society* (e.g. if the technology benefits the developer but creates negative externalities for society, for example, by creating undue risk or undermining valuable institutions);
6. Favour *society* at the expense of the *user* (e.g. if the technology unduly limits user-freedom for the sake of a collective goal such as national security);
7. Harm the *user*, simpliciter (e.g. if the technology breaks in a way that harms the user without benefiting anyone else);
8. Harm *society*, simpliciter (e.g. if the technology is unfair or has destructive social consequences without benefiting anyone else).

In many cases, an intuitive understanding of proportionality may be sufficient to detect whether an advanced AI assistant has erred and ceased to be sufficiently value-aligned. However, understood on a more foundational level, we have suggested that the notion of proportionality itself needs to be understood by reference to wider societal principles or ideals, including views about justice and about civil and human rights.

What implications does the preceding analysis and account of value alignment have for the development, design and release of advanced AI assistants of the kind that form the central focus of this paper? First, it points towards the need for a more nuanced understanding of the *harms* and *ideals* that underpin the positive development of this technology. The notion that advanced AI assistants should be helpful, honest and harmless is a useful starting point. However, we also need a more complete understanding of how these values apply to different contexts, and of various failure modes and mitigation techniques. In particular, we need to consider who this technology has the potential to harm, in what way this individual or group might be harmed, and whether the nature of the harm varies for different parties or according to different contexts. By exploring the intersection between advanced AI assistants and *Well-being*, *Safety*, *Privacy*, *Trust*, *Malicious Uses*, *Misinformation*, *Anthropomorphism*, *Manipulation and Persuasion*, *Appropriate Relationships, Cooperation*, *Equity and Access*, *Economic Impact* and *Environmental Impact*, we hope to develop a more complete understanding of these questions.

Second, developers need to take an *inclusive* view of value alignment and should not focus on user *preferences* or responsiveness to user *intentions* alone (see Chapter 19). After all, there may be a disconnect between individual preferences and what is good for the user (see Chapter 6). There may also be a disconnect between the preferences of raters used to train AI assistants and what is good for society. Moreover, deference to user intentions must be bounded in certain ways. This is implicitly recognised when developers train models to respect constraints or rules (Bai et al., 2022b; Glaese et al., 2022). However, deeper analysis, monitoring and evaluation is needed – both at the level of the user and society – to ensure that appropriate safeguards and constraints operate across a full range of contexts (see Chapter 19).

Finally, it will be fruitful to continue to explore ways of developing and training AI assistants that are consonant with *democratic principles* and *value pluralism*. If advanced AI assistants turn out to be a powerful and pervasive technology that plays an important role in many people's lives, then the question of justification

and legitimacy will not go away. Rather, a recurrent question will be: Who gave you the right to decide? By exploring mechanisms that enable more participatory and democratic value elicitation, model training and evaluation, it may be possible to create artefacts that complement prevailing societal ideals and social practices by supporting them in relevant ways and by garnering the right kind of public support.

Chapter 6

Well-being

Nenad Tomašev, Ira Ktena, Arianna Manzini, Geoff Keeling, Zeb Kurth-Nelson, Andrew Barakat, John Oliver Siy, Iason Gabriel

Synopsis: We build on theoretical and empirical literature on *conceptualisations* and *measurements* of human well-being from philosophy, psychology, health and social sciences to discuss how advanced AI assistants should be designed and developed to *align with user well-being*. We identify key technical and normative challenges around the understanding of well-being that AI assistants should align with, the data and proxies that should be used to appropriately model user well-being, and the role that user preferences should play in designing well-being-centred AI assistants. The complexity surrounding human well-being requires the design of AI assistants to be informed by domain experts across different AI application domains and rooted in lived experience.

6.1. Introduction

Narratives surrounding the introduction of new technologies often emphasise improving productivity and off-loading unpleasant tasks to free up human time for enjoyable activities. Arguably, the current technology is yet to fully deliver on that promise (Wajcman, 2020). Furthermore, in recent years, studies of mobile phone use and social media have suggested a more challenging reality of rising online toxicity (Haidt and Schmidt, 2023), which has at times resulted in real-world physical harms (Lomas, 2022). These unanticipated adverse outcomes have sparked a debate around the effects of technological advances on human well-being more widely.

Interactions with AI assistants are already beginning to permeate a wide range of domains in users' daily lives. One need only look at the rapid pace of adoption of publicly accessible large language models (LLMs) such as ChatGPT, and the numerous applications that are currently being developed and powered by this kind of technology, to understand the scale of these potential effects. The new capabilities enabled by advanced AI assistants present us with the opportunity and responsibility to re-evaluate and reimagine our relationship with technology, so that it is utilised to support and facilitate human well-being (McGillivray, 2007) and flourishing (see also Chapter 11). However, ensuring alignment between developers' intentions, AI systems' behaviour and users' well-being comes with numerous challenges (van der Maden et al., 2023; Xiang, 2023). Thus, in this chapter we aim to investigate how we can develop AI assistants that are *aligned with user well-being*.

An important consideration for value alignment, in the context of well-being alignment, is whether technology should only avoid reducing well-being or should actively improve it (see Chapter 5). While we believe that AI assistants *should not harm* user well-being, we remain agnostic about whether they should be developed with the overall aim of elevating well-being above a baseline level across the board (Gable and Haidt, 2005).

We also note at the outset that this chapter discusses well-being in relation to *users* of AI assistants. It leaves considerations about the implications for non-users to other chapters (see Chapter 15).

This section is structured as follows. First, we review conceptualisations of human well-being by drawing on an extensive theoretical and empirical literature from philosophy, social sciences, psychology and health. We then discuss the challenges associated with research efforts to measure well-being, as an essential step in understanding the causes and consequences of human flourishing, and devising effective policy interventions for supporting and improving well-being. Third, we discuss the different ways in which existing technologies have influenced user well-being. We draw inspiration from these ideas and related learning to outline the opportunities and risks that arise from the design, development and deployment of well-being-centred AI assistants. We conclude by providing actionable recommendations aimed at developers of advanced AI assistants.

6.2. Understanding Well-being

The philosophy of well-being

Although well-being is a foundational aspect of human experience, it is notoriously difficult to formalise. Across disciplines, well-being has been qualified to include physical and mental health (Levin, 2020), engagement, optimism, self-esteem, experiencing positive emotions like happiness, contentment and overall life satisfaction (Huppert, 2009), finding meaning and purpose, leading a life of virtue and forming and maintaining close social relationships with other people (VanderWeele, 2017).

A helpful starting point for systematising this large body of knowledge is provided by the philosophy of well-being, which focuses on what is *intrinsically* good (i.e. valuable in itself) for human beings, as opposed to what is *instrumentally* good (as a means to another end) for us. Three main theories of well-being can be distinguished in this space.

Hedonism, which is associated with classical utilitarian philosophers like Jeremy Bentham (Bentham, 1970) and Francis Edgeworth (Edgeworth, 1879), alongside the British Empiricists such as Thomas Hobbes (Hobbes, 1994), David Hume (Hume, 1998) and John Stuart Mill (Bentham and Mill, 2004), equates well-being to the balance of pleasure (or happiness) over pain (or suffering). According to hedonism, facts about what is good for an individual depend only on facts about their pleasure and pain. What makes it the case that substantive goods such as friendship, completing a university degree and winning the lottery are *good for* particular people is precisely that these goods have the property of increasing their pleasure or decreasing their pain.[1] Hedonism has been criticised for being reductionist, in that it considers pleasure as the only non-instrumental good, but presumably other non-instrumental goods exist, such as those associated with meaningful accomplishment (Nozick, 1974). Hedonism has also been criticised for treating all forms of pleasure (physical ones, intellectual ones and even 'evil pleasures' (Crisp, 2011)) as equally significant. In addition, hedonism has been criticised on the grounds that it is not obvious how to measure the quality of subjective experiences in a way that allows for interpersonal comparisons in well-being (Fisher, 2007; Robbins, 1938; Wicksteed, 1910). However, such objections at best target hedonism as a guide for practical decisions in economics and public policy, and they have less obvious relevance to the plausibility of hedonism as a theory of well-being.[2]

[1] Some proponents of hedonism reject the language of pleasure and pain. Consider Roger Crisp (Crisp, 2006): '[W]e should try as far as possible to avoid talk of "pleasure", for a reason noted by Aristotle and many writers since: "[T]he bodily pleasures have taken possession of the name because it is those that people steer for most often, and all share in them". This, of course, is why a version of the philosophy of swine objection against hedonism – that the hedonist is advocating the life of sensualism – arises so readily. To avoid such difficulties, let me use "enjoyment" instead of "pleasure", and "suffering" instead of "pain."'

[2] For a recent philosophical treatment of the measurement-theoretic questions presented by hedonism and forms of utilitarianism that

Desire theories of well-being overcome some of these objections by defining well-being as the fulfilment of one's preferences or desires. However, disagreement exists around what conceptualisation of 'preferences' should be considered as constitutive of well-being (see also Chapter 5). Some desire theorists focus on the satisfaction of preferences that are about what one wants their life to be like overall, or about the shape and content they desire their life to have, as opposed to immediate short-term desires and wishes. Another important dispute concerns accounts that define well-being as the satisfaction of 'ideal' preferences (i.e. those that one would have if they were fully informed and had time to deliberate clearly and rationally on their wishes), as opposed to those preferences that are simply revealed through one's behaviour (Otsuka, 2015). Across these variations, a key feature of desire theories is that a person needs to *desire* a particular good or valuable thing for that thing to contribute to their well-being. This implies that it is ultimately up to each individual to decide what makes their life go well for them (Parfit, 1984). This view may, however, fail to accommodate situations where it is the unanticipated and surprising things that make people feel good, things that they did not necessarily know about or appreciate beforehand.

Finally, *objective list theories* hold that well-being consists in a list of objectively valuable things (e.g. pleasure, knowledge and deep relationships), regardless of how we individually feel about them (Parfit, 1984; Ryan et al., 2013). Objective list theories run into the challenge of having to answer complex questions such as what goes on the well-being list, who gets to – or should – make that decision, and whether it is morally permissible to disregard an individual's preferences as misguided if they do not match what is objectively valuable or important (Parfit, 1984).

The science of well-being

Moving beyond philosophy, there have been numerous pieces of theoretical and empirical research in psychology, health and across the social sciences aimed at elucidating the underlying *drivers* that contribute to experiences of well-being (and lack thereof) (Das et al., 2020a; Helliwell and Aknin, 2018). For example, self-determination theory identifies autonomy, competence and relatedness (Ryan et al., 2013) as key factors that, when neglected, may be detrimental to one's sense of well-being. Engaging in creative (Marshall et al., 2014) and artistic (Tay et al., 2018) work, goal fulfilment (Steca et al., 2016) and social presence (Chang and Hsu, 2016) have also been shown to have similarly positive effects.

Conceptions of human flourishing and human dignity also play a pivotal role in wider research focused on the relationship between well-being and social inequality, distributions of power, and human rights (Kleinig and Evans, 2013). In this context, a concern for individual well-being intersects with other priorities, such as the reduction of poverty and disease (Rao and Min, 2018) and the promotion of freedom and justice (Dolan and White, 2007; Sen, 2001), as part of efforts aimed at developing effective interventions for improving well-being at the societal level. This strand of research has drawn further attention to the *contextual nature* of well-being (see also Chapters 5 and 11), which is influenced by broader socioecological (King et al., 2014), economic (Summers et al., 2014), political, cultural (Diener et al., 2018) and environmental factors which are often beyond individual control (Docherty and Biega, 2022). Sustainable approaches to improving overall well-being (Collste et al., 2021; Holdren, 2008) may involve trade-offs between short-term and long-term societal needs. They would therefore require reaching a wider consensus among policymakers and society at large.

Among other things, the above considerations illustrate the considerable *complexity* that surrounds conceptualisations and experiences of well-being. Given this complexity, developers that endeavour to build AI

employ a hedonistic axiology, see Narens and Skyrms (2020).

assistants that enhance well-being must be clear and transparent about how their *underlying assumptions* about human well-being inform the design of these technologies.

6.3. Measuring Well-being

In science and politics, measuring well-being is considered essential for understanding the *causes* and *consequences* of human flourishing, and for devising and evaluating *interventions* that can help people live a good life (Alexandrova, 2017). In this section, we discuss various approaches to measuring well-being and the underlying conceptions that give rise to them. Given that such measurements will be necessary for aligning AI agents with the well-being of their users and the challenges that arise from the non-observable nature of some of its facets, we present proxies that have been used in practice. Finally, we discuss the importance of distinguishing between causal links and plain associations to identify effective interventions that positively influence well-being.

Approaches to well-being measurement

There are two main approaches to well-being measurement (Voukelatou et al., 2021). The *subjective approach* studies people's subjective evaluation of the quality of their own life, through methods such as self-report questionnaires that inquire about life satisfaction or happiness. In contrast, efforts to operationalise a more *objective conception* of well-being tend to measure observable dimensions of a fulfilling life via indicators such as education, household income or consumption expenditure. The purported objective well-being indicators are often linked to and derived from the subjective well-being measures and outcomes, and there is ongoing research aimed at bridging the gap between them (Layard and De Neve, 2023).

Both approaches have important limitations for comparisons between different *cultural or demographic* groups (Heine et al., 2002; Krueger and Schkade, 2008; Krueger and Stone, 2014). Similar responses to self-reported questionnaires can fail to capture underlying differences between groups in their experiences, as well as the underlying drivers of well-being, their values, desires and preferences. Cultural, demographic and individual differences in the interpretation of questions and response scales may also play a role in how otherwise equivalent well-being experiences are reported as distinct by different individuals, thus resulting in reporting bias (Krueger and Stone, 2014). At the same time, different groups place different values on observable dimensions like consumption expenditure, meaning that those dimensions measure what actually matters to some groups more than others. These two intertwined issues highlight the need for careful considerations about what well-being-aligned personal assistants should optimise for when they are deployed across cultures and demographics to avoid the risk of harming certain groups or disadvantaging them by failing to best meet their needs.

Underlying conceptions of well-being and their role in informing measurement

Any well-being metric comes with underlying assumptions about *what human well-being is* (see Section 6.2). For example, social psychology researchers have proposed the U-index, which is aimed at measuring the average amount of time a person spends in an unpleasant state (Kahneman and Krueger, 2006). The idea behind this index is that most practical interventions are aimed at reducing suffering or unpleasant emotional states rather than maximising happiness, and this should be reflected in metrics used to measure well-being. There is also a tendency for indices to focus on the satisfaction of basic human needs (Smith et al., 2013) (e.g. basic economic indicators and health), which are seen as prerequisite for a deeper and more holistic sense of well-being, and yet are still unreachable to many people in today's societies. This hierarchy of needs has been

questioned, especially in terms of them needing to be met sequentially (Rojas and Guardiola, 2016; Rojas et al., 2023). Critics have countered that the alleviation of suffering or satisfaction of basic human needs fall short of accounting for *eudaimonic aspects* of well-being (Dolan and Metcalfe, 2011; Kapteyn et al., 2015),[3] so they are insufficient for establishing that a person is flourishing or that their life is going well. This debate raises the question of what conception of user well-being AI assistants should align with – whether they should help us to just meet our very basic needs or also identify ways in which we can flourish and lead fulfilling lives (see Chapter 5).

Well-being proxies

Developing aligned AI assistants requires not only a conception of user well-being but also *practical metrics*. For most of the definitions of well-being outlined above, well-being itself is not directly observable. For example, there is no tool for measuring hedonic pleasure. Researchers must therefore identify measurable *proxies* – like self-report about hedonic pleasure – that are expected to correlate with well-being. A good proxy has high *construct validity*, meaning that it relates closely to the unobservable construct of well-being (Alexandrova, 2017). However, there is often a trade-off between construct validity and ease of measurement. For example, financial welfare may be relatively easy to measure, but it often fails to translate into life satisfaction (Layard, 2010; see also Chapter 19), and as a proxy for well-being, it has poor construct validity. Governments and researchers have developed many kinds of metrics for assessing well-being, each of which has strengths and weaknesses.[4]

Causality

Finally, a fundamental concern in well-being measurement is that most existing metrics of well-being rely on correlates (Smith et al., 2013). The absence of *causal understanding* of the link between underlying determinants and well-being is a major obstacle for developing reliable and effective interventions to support human flourishing, and for the design of AI assistants that align with user well-being. Interventions that are designed based on associations derived from retrospective observations may fail when deployed in practice, especially in a rapidly evolving real-world environment where such associations may easily break unless they represent verified causal links. These links may be hard to establish from retrospective data alone, thus necessitating experimentation and learning from outcomes of targeted interventions (Walton and Wilson, 2018; Wilson, 2011).

6.4. Influence of Current Technology on Well-being

The well-being-centred design of AI assistants can be informed by other technologies. Indeed, well-being, satisfaction and sustainability measures are starting to be integrated into product development and product

[3] Eudaimonic well-being refers to subjective experiences associated with living a virtuous life in pursuit of human excellence (Niemiec, 2014). Eudaimonic well-being definitions may vary, but they tend to include aspects of meaning, value and relevance to a broader context, personal growth, self-realisation and maturity, excellence, ethics, authenticity, and autonomy (Huta, 2015; Huta and Waterman, 2014).

[4] Smith et al. (Smith et al., 2013) provide a comprehensive overview of such indices, including the Quality of Life (QOL) Index for Developed Countries (Diener, 1995), Australian Unity Well-being Index (Cummins et al., 2003), Happy Planet Index 2.0 (Marks, 2006), Hong Kong QOL 2008 (Chan et al., 2005), Human Development Index (UNDP, 1990), Sustainable Society Index (Van de Kerk and Manuel, 2008), Index of Child Well-being in Europe (Bradshaw and Richardson, 2009), The Economist Intelligence Unit's QOL Index (Economist Intelligence Unit, 2005), Child and Youth Well-being Index (Land et al., 2001), Nova Scotia 2008 GPI (Pannozzo, 2009), The State of the Commonwealth Index (Watts, 2004), Fordham Index of Social Health (Miringoff and Miringoff, 1999), National Well-being: Life Satisfaction (Vemuri and Costanza, 2006), The Well-being of Nations (Prescott-Allen, 2001), OECD Better Life Initiative (OECD, 2011), Gallup Healthways Well-being Index (Gallup-Healthways, 2009), QOL 2007 in Twelve of New Zealand's Cities (Jamieson, 2007), Well-being in EU Countries Multidimensional Index of Sustainability (Distaso, 2007) and Gross National Happiness (Ura, 2008).

assessment (Kramer et al., 2014; Stray, 2020; Wen et al., 2016), which has revealed a range of effects. Increased screen time and addictive content have been shown by some research to have negative consequences (Orben and Przybylski, 2019), including feelings of loneliness (Wilson, 2018). At the same time, e-health technologies are being designed in the hope of improving health outcomes (Cechetti et al., 2019; Granja et al., 2018; Hors-Fraile et al., 2018) and the efficacy of mental health care (Blandford, 2019; Eysenbach et al., 2001).

Taken together, prior studies indicate that there may be a significant opportunity for technology to help address well-being issues and steer people towards a healthier life. There is some urgency, given the prevalence of stress, poor sleep quality, insufficient exercise and unhealthy eating resulting in long-term chronic health issues impacting both individual well-being and society at large (Abe and Abe, 2019; Abegunde et al., 2007; Hargens et al., 2013; Mascie-Taylor and Karim, 2003). Personalised interventions using techniques such as mindfulness and meditation have shown promise (Isaacs et al., 2013; Jeong and Breazeal, 2016), but the evidence is still mixed (Gál et al., 2021; Howells et al., 2016). In particular, digital well-being initiatives may fail to deliver lasting impact due to difficulties in influencing the formation and reinforcement of new habits (Monge Roffarello and De Russis, 2019). Digital interventions should therefore build on the existing literature on promoting habit formation (Bandura, 2013; Eyal, 2014; Gardner et al., 2023; Lally and Gardner, 2013), and there is an opportunity for AI in particular to play a role in reframing and optimising behavioural interventions to make them easier to follow and more satisfying for the users.

As one key example, *recommender systems* remain at the centre of intense scholarly debate regarding their potential for behavioural adjustment through tailored recommendations and whether their benefits outweigh their risks (Chen et al., 2023a; Przybylski and Weinstein, 2017; Stray et al., 2022). To make targeted and effective recommendations in the moment, and get positive signals from users' interactions with the system, recommender systems tend to optimise for *short-term reward* rather than an accumulated value of recommendations over a longer time span (Burr et al., 2018). While there may be cases where short-term optimisation is not intrinsically misaligned with longer-term goals and well-being, one does not imply the other, so a level of caution is required. Different apps and products compete for attention, making it harder to implement sustainable, healthy incentives to promote users' well-being due to the resulting fragmentation of attention and exposure to more addictive app surfaces (Bhargava and Velasquez, 2021). However, recommender systems could potentially help us to plan our daily activities (Khwaja et al., 2019), promote healthy choices (Hors-Fraile et al., 2018), improve our nutrition (Toledo et al., 2019) and tailor our lifestyles (Hammer et al., 2015) when designed with happiness and well-being in mind (Gyrard and Sheth, 2020; Nouh et al., 2019). Thus, considerations for aligning recommender systems with user well-being have been made (Stray et al., 2021) in line with measures designed to address the broader value alignment problem (see Chapter 5). Proposed steps in aligning recommender systems involve identifying the most important outcomes, operationalising these concerns via hand-crafted or learned metrics, and utilising these metrics to adjust recommendation behaviour. This is part of a broader and increasing appreciation of ways in which technological systems may be purposeful and *support well-being and human potential* (Calvo and Peters, 2014). Through this endeavour, various types of product specifications that focus on different aspects of user well-being – including pragmatic, hedonic and eudaimonic dimensions – have been developed (Kamp and Desmet, 2014).

Given the cultural, social, ethical and psychological variables that influence well-being, it becomes evident that *partnering with social scientists* is vital to the success of these technologies in grounding the design decisions and efforts in scientific research. Developing cross-disciplinary theory through these partnerships is also essential to weaving the formalisms of these different fields together. Such partnerships could aim to empower and involve psychologists and social scientists in early stages of AI system design, and place them in a leading role in the design of well-being principles and metrics.

6.5. Opportunities and Risks with AI Assistants

It is quite possible that interactions with AI assistants will soon permeate many areas of our daily lives, and that these systems will develop a deeply personalised understanding of users' needs and preferences. It has, therefore, become urgent to ensure that such information is not used in ways that diminish user well-being, but rather that it is used to support or even enhance it. This requires that a range of open challenges be addressed (see also Chapter 13).

Well-being data collection

Unsurprisingly, conversational agents are increasingly considered as an assistant paradigm for delivering targeted interventions for improving physical and mental health (Kimani et al., 2019; Kocaballi et al., 2020). Future AI assistants may be interact with their users via various digital surfaces or alternatively they could be situated within affective social robots (Wairagkar et al., 2021). In both cases, to help their optimisation for well-being, AI assistants may need to *obtain data* about their users involving subjective metrics of well-being (e.g. how well they think their needs are being met, their progress towards their goals and their overall feeling of fulfilment and satisfaction) or to collect data about more objective indices (e.g. income and personal health). This data could come from multiple sources such as multi-modal sensory monitoring (Fahim et al., 2014; Lane et al., 2014), mobile data (Bogomolov et al., 2013), wearables (Nahavandi et al., 2022), conversational signals or self-reported happiness levels.

Despite the existence of clear opportunities arising from rich, integrated data, there are fundamental concerns about the efforts to comprehensively model well-being and capture everything that really matters. As explained above, using poor proxies runs the risk of misrepresenting well-being and, hence, of steering users away from achieving happiness and a flourishing life.[5] Moreover, any such data collection and integration comes with a set of *privacy concerns* (see Chapter 13). It may be possible to at least partially mitigate some of the highlighted issues by including a rich set of proxies to begin with, updating these proxies through participatory and democratic mechanisms, implementing detailed monitoring of proxy outcomes, avoiding over-fitting onto any such simplified objective and observing best ethical practice in the field of data collection.

User preferences integration

One of the most important open questions concerns the role that *user preferences* should play in designing well-being-centred AI assistants. This choice could potentially lead to greater *user agency and autonomy* – qualities which also feature in more objective conceptualisations of well-being. However, for this to be the case, the desires which AI assistants help to advance must be of the right *kind* (Mitelut et al., 2023). Thus, appropriately aligned AI assistants should be designed to understand user desires and motivations. However, current approaches to integrating user preferences into AI systems' decision-making encounter challenges and limitations that risk *undermining* the well-being related goals that they might otherwise serve (see also Chapter 5).

Technical challenges

Some of these challenges are technical. Using human preferences in reinforcement learning is a research programme with a long history (Bai et al., 2022a; Christiano et al., 2017; Jaques et al., 2019; Wirth et al.,

[5] For example, it has been hypothesised that health, measured through the lens of current morbidity, ought to be a strong proxy for personal well-being towards the end of life, as it underpins the ability to meet other well-being objectives. Yet a study (Gerlach et al., 2017) established that morbidity accounts for only 20% of the observed variance in reported well-being in that population.

2017), and the development of methods to align language models with human preferences continues to be an active area of research (Go et al., 2023). More recently, *reinforcement learning from human feedback* has been a key component in shaping the behaviour of conversational (Bai et al., 2022a; Ouyang et al., 2022) and multimodal interactive (Abramson et al., 2022) agents (see Chapter 3).

In fact, even when preferences are not explicitly given, they can often be derived implicitly from contextual cues, user behaviour or prior interaction histories (Holland et al., 2003; Liu et al., 2017). However, the explicit preferences, and those inferred via observational learning, may at times *conflict*, meaning that real-world interactive scenarios necessitate frameworks for dealing with such inconsistencies (Oguego et al., 2018). Assistive agents also need to integrate user preferences into *planning* (Benton et al., 2012; Jorge et al., 2008) to ensure adherence when executing tasks for the users. Explicit planning with language models is an active area of research, with some promising directions involving tree-search or graph-based reasoning (Besta et al., 2023; Long, 2023). Yet, when planning, well-being-driven AI assistants may still need to optimise towards multiple objectives simultaneously (Hayes et al., 2022; Yang et al., 2019) to find effective ways of navigating between them.

To address this complexity, AI assistants could, for example, generate a diverse set of possible plans, recommendations and outcomes (Nguyen et al., 2012) for users to choose from, each of which evidence different trade-offs between various personal goals. Providing users with agency in these decisions may well be a critical step towards human empowerment in their interactions with AI assistive technology. Nevertheless, achieving quality and diversity in plans provided by AI agents remains an open problem and is an active area of research (Lim et al., 2022; Zahavy et al., 2021, 2022, 2023).

Normative challenges

There are also deeper normative challenges about whether preference satisfaction is the central goal AI assistants should aim for, even when the goal is to align agent behaviour with user well-being (see Chapter 5). Individual preferences may at times be irrational or inconsistent with each other (Gabriel, 2020). This raises the question of what assistants should do about these inconsistencies, and the possible nuances associated with them, to avoid replicating the limitations that come with aggregated rater preferences in training language models (Gordon et al., 2022; see Chapter 3). Individual preferences may also be in conflict with what will make a user *flourish*. Some users may not have strongly held preferences, and some may not have the language to communicate them accurately. Assistants that align with a theory of well-being based on desire satisfaction may be more easily directed towards spurious objectives than the underlying objective list account. Additionally, if AI assistants only have access to users' *revealed* rather than *ideal* preferences, they may end up satisfying their immediate and short-term goals at the expense of long-term well-being (Burr et al., 2018). This is particularly likely to happen in contexts where there are commercial incentives for focusing on *short-term user gratification*, and therefore for developing products that users *like* and *use* over those that promote their overall and long-term well-being. Identifying a way to balance users' short-term wishes with their long-term well-being goals remains one of the most important open questions in the design of AI assistants, and it is something that future research will have to address. Ultimately, AI assistants will need to find ways of managing these trade-offs and supporting users while also respecting their stated preferences and wishes (see Chapter 5).

Risks of co-adaptation and manipulation

As AI assistants become more integrated into users' daily lives through repeated interactions, it is also important to consider how existing user preferences may start to be shaped and how new ones may come about (Liang, 2019). There is a non-negligible risk that, through current interactions, AI assistants may *influence future* user

preferences in ways that create ethical challenges (Ashton and Franklin, 2022). For example, in increasingly deep user–assistant relationships, it may become hard to distinguish instances where systems have clearly improved user well-being from cases in which users have adapted their behaviour to that of the assistant in ways which may sometimes invalidate the usefulness of the originally designed well-being metrics.[6] This risk is particularly salient in the case of recommender systems due to the emergence of degenerate feedback loops (Jiang et al., 2019). In the case of advanced AI assistants, the concern is that the AI system could, intentionally or unintentionally, *influence and steer* user behaviours in unanticipated ways, and this may potentially obscure some well-being issues if the corresponding metrics were not designed robustly, since subtle behavioural shifts may not always be easy to identify. For example, we may at times mistake manipulative behaviours for helpful behaviours that contribute to user well-being (see Chapter 9). This could be the result of undesirable system behaviours like sycophancy (Perez et al., 2022b), reward hacking (Hadfield-Menell et al., 2017), reward misidentifiaction and causal confusion (Tien et al., 2023) in preference-based learning (see Chapter 7). Thus, ensuring ethical deployment of well-being-centred AI assistants will require advances in our existing frameworks for robustly inspecting and verifying agent behaviour (see Chapter 19).

6.6. Outlook

Despite these concerns and challenges, we believe that there is untapped *potential* in aiming for the development of digital personal assistants that are able to support and improve individual physical and mental well-being (Balasubramanian et al., 2021; Grossman et al., 2004; Gu et al., 2015). Digital personal assistants could facilitate these improvements in user well-being either by directly optimising for well-being outcomes or as a secondary outcome following improvements in other areas – for example, improved problem-solving and planning abilities. Promising avenues for future research and assistant development in this area include: the ability of LLMs to learn and adopt behavioural rules and principles (Bai et al., 2022b), to simulate human behaviour (Park et al., 2023a) and to rapidly personalise content based on prior similar interactions (Welch et al., 2022), as well as the simplicity with which aligned recommendations can be elicited through instructions (Zhang et al., 2023).[7] Embedding humanistic principles in AI assistants may not only help align the technology with users and society but also help those who interact with assistants better realise their own aspirations by finding meaning and support in an ever-changing world riddled with challenges (see Chapters 17 and 18). As Lehman and others have proposed (Alberts et al., 2024; Fromm, 2000; Lehman, 2023), for this to be achieved, AI assistants might need to display qualities that are analogous to *deep care* for people, *responsibility* for assisting positive actions, *respect* for the ways in which people wish to develop – and how they wish to go about achieving their goals – and holistic *understanding* of people's needs. This is an *aspirational vision*, necessitating deep syntheses across fields, but one towards which we can hope to orient the field and start making meaningful progress.

6.7. Conclusion

Understanding, measuring and intervening to better support human well-being has been the goal of long-standing research efforts across disciplines like philosophy, psychology, public health and the social sciences, from which we can learn to design AI assistants that align with user well-being. We conclude with a list of recommendations that technologists developing these systems may want to consider.

[6] In a related example, a study that compared the performance of a search system with several of its intentionally degraded versions found that, somewhat counter-intuitively, the ultimate success rate of the degraded system versions was just as high as the original one (Smith and Kantor, 2008). However, user behaviour was measurably different, because users had developed strategies to compensate for system weaknesses.

[7] However, see Chapter 5 for an analysis of the limitations.

1. **Seek deeper community involvement and empowerment of domain experts**: Domain experts across fields of human psychology, health and social sciences have expertise in understanding and measuring well-being. There is, therefore, a need for deeper involvement and empowerment of these experts in AI assistant design and development (Peters et al., 2018). Given the complexity and diversity of experiences of well-being, it is also critical to employ participatory approaches for different demographic and cultural groups to inform the design of these technologies (Martin Jr et al., 2020).
2. **Adopt a clear and context-dependent understanding of well-being**: The complex and multifaceted nature of human well-being requires developers to be clear and transparent about what conceptualisation of human well-being informs their design decisions when developing AI assistants. To ensure fair distribution of benefits across groups, AI assistant design should avoid the pitfall of trying to impose a universalist conception. It should instead accommodate cross-cultural and demographic differences in subjective and objective perceptions of well-being (McGregor, 2018).
3. **Identify and use appropriate proxies:** One common assumption in discourses around technological advances is that, by improving efficiency in executing tasks, and so overall productivity, new technologies will have a positive effect on people's well-being by default. This view considers a single facet of well-being while disregarding its potential negative impact on other aspects that drive human flourishing. Developing AI assistants that do not harm, but rather support or enhance, human well-being requires technologists to identify and use appropriate proxies by leveraging empirical studies on well-being measurement.
4. **Understand the complexity of user preferences**: To align with users' well-being, AI assistants need to understand user goals and preferences and proactively solicit and integrate their feedback. While most existing technologies optimise for providing short-term value, well-being-centred AI assistants will need to differentiate between short-term and long-term preferences, as well as ideal and revealed preferences, and avoid over-indexing on immediate preference satisfaction.
5. **Ensure effective and ethical data collection**: Empirical research is required to ensure that identified proxies serve the purpose of supporting user well-being. Here researchers need to consider plausible methods for enabling AI assistants to collect subjective and objective well-being data from and about users. Research efforts should focus on methods to integrate such rich data to model user well-being appropriately while ensuring user privacy is respected.
6. **Monitoring**: Even for AI assistants designed with well-being principles in mind, it is important to incorporate a number of explicit metrics to help evaluate the consequences of their use at deployment. These metrics should be developed and selected by domain experts working closely with the technical teams. Metric design may also inform data collection practices as much as data availability may itself inform the feasibility of implementing certain metrics.

There is a tangible need for interdisciplinary research to come together, in an inclusive and participatory manner, to help inform ways in which current and future assistive AI technology may help to address well-being needs, and thereby shape policies and governance for ethical design (Feijóo et al., 2020). If appropriately designed, developed and deployed, advanced AI assistants have the potential to improve user well-being and may play an important positive role in our lives. Yet, given the role that people, community and social connections play in our overall well-being, much of the positive impact of future AI assistants may come not from our direct interactions with them but from the way in which they enable us to foster and strengthen our social bonds with others (see Chapter 11).

Chapter 7

Safety

Zachary Kenton, Victoria Krakovna, Verena Rieser, Geoff Keeling, Iason Gabriel

Synopsis: This chapter focuses on *dangerous situations* that may arise in the context of AI assistant systems, with a particular emphasis on the *safety* of advanced AI assistants. It begins by providing some background information about *safety engineering* and safety in the context of AI. The chapter then explores some *concrete examples* of harms involving recent assistants based on large language models (LLMs). Building on this foundation, it then considers safety for advanced AI assistants by looking at some hypothetical harms and investigating two possible drivers of these outcomes: *capability failures* and *goal-related failures*. The chapter concludes by exploring *mitigation techniques* for safety risk and avenues for future research.

7.1. Introduction

AI safety is a broad topic concerned with *mitigating risks* and *minimising harms* that arise from the development and deployment of AI. In this context, *harms* are bad outcomes that actually occur, for example death or human suffering, whereas *risk* refers to the probability of the harm occurring. Moreover, while the field of AI ethics addresses a number of the risks posed by AI systems, the field of AI safety focuses primarily on a set of *serious* and *relatively direct* risks involving harms such as the real and significant chance of death, physical injury and psychological damage (e.g. through abuse, blackmail and coercion, as well as property damage and theft).[1] Of particular importance for AI safety research are the risks and harms that are possible in today's cutting-edge AI systems and those which are amplified when the AI system has more powerful capabilities (Anwar et al., 2024; Hendrycks et al., 2022; Phuong et al., 2024).

Risks from AI can come in various forms, but we can categorise them as follows:[2]

- *Accident risks*, which arise when AI systems do something different from what their designers intended.
- *Misuse risks*, which arise through misuse that is either unintentional or caused by malicious actors.
- *Structural risks*, which are unintended bad outcomes that occur despite the AI doing what the designers intended it to do in a more proximate sense.

[1] There have been efforts to broaden the scope of safety in recent years – and these problems undoubtedly warrant attention in their own right (Bender et al., 2021; Dinan et al., 2022; Shelby et al., 2023; Weidinger et al., 2022b).

[2] This categorisation does not necessarily partition the space of risks, but rather intends to be a useful practical guide that should help in many situations when trying to think about a typical risk.

In this chapter on safety, we focus primarily on *accidents*, as malicious use and structural risks arising from the development and deployment are largely covered elsewhere in this paper (see Chapters 8, 14, 15, 16, 17 and 18). It should also be noted that accident risks (and misuse risks) could potentially have large society-scale effects (i.e. of a similar magnitude to structural risks that arise in the context of inequality or automation and unemployment).

We next discuss safety in the context of engineering before looking at these considerations more specifically in the context of AI, with a focus on harms from recent LLM-based assistants. Building on this foundation, we then explore two kinds of failure that may affect the safety of advanced AI assistants: capability failures and goal-related failures. Finally, we conclude with a discussion of mitigation techniques and avenues for future research.

7.2. Safety Engineering

In the context of *engineering*, safety is aimed at ensuring that engineered systems provide acceptable levels of safety in settings where there is potential for harm (generally assumed to be physical and life-critical), even in the face of the failure of system components.

Some safety engineering methodologies that are designed to identify and address undesired outcomes early in the development process have been considered in the context of AI systems. For example, Rismani et al. (2022) consider the applicability of failure mode and effects analysis (FMEA) and system theoretic process analysis (STPA) to this context. FMEA takes a fairly reductive (divide-and-conquer) approach to identifying failure over the development life cycle (Carlson, 2012). For each *component* of the system, FMEA considers the possible *failure modes*, their *severity*, *likelihood* and *chance of detection* before assigning a risk priority number from which prioritisation can occur. Unlike FMEA's reductive approach, STPA focuses on emergent phenomena based on *interactions* between components (rather than just the components themselves) and *feedback loops* between the engineered system and the wider system within which it is embedded. It aims to model the full sociotechnical system using multiple controller feedback loops. It then uses that model to identify unsafe control actions. By surveying machine-learning (ML) researchers, Rismani et al. (2022) find that FMEA and STPA could in principle be helpful for risk assessment, but it is unclear whether these have actually been used in practice so far.

Normal accident theory (Perrow, 1999) argues that, due to the complexity of our society's systems, multiple and unexpected failures are fundamental and that accidents are unavoidable. It has been applied in multiple engineering domains such as aerospace and nuclear systems. Maas (2018) argues that AI systems, including narrow AI applications, are also prone to normal accident theory failures due to their *complexity* and *opaqueness*, their *interaction speed*, the multiple *competing objectives* of their designers (safety being only one objective) and the competitive *race dynamics* (see also Bianchi et al., 2023). This suggests interventions on the *levers* of normal accident risk are important. They include policies encouraging *explainable/interpretable* AI (see e.g. Räuker et al., 2023); *limiting integration* into societally important functions and *restricting* AI *autonomy* and/or *speed* of interactions (see e.g. Christiano); clarification and enforcement to intended operational *domains*; and better safety and ethical *training* of ML practitioners, including sharing safety expertise between organisations (Ho et al., 2023).

7.3. AI Safety

Background

While some aspects of general software safety engineering are applicable to AI, we cannot rely solely on those methodologies for creating safe AI systems (Hendrycks et al., 2022). Software safety engineering approaches rely on the underlying engineered system having a control structure that is *explicitly* programmed by humans. AI control structures are instead *learnt* (in the context of ML, which is our focus here) via optimisation and stored in inscrutable weights (in the case of large-scale deep learning, which is our focus). This ML control structure is therefore difficult to assess for *completeness* and *coverage*; is *fragile* and the fixes are *complicated*; involves *non-modularity* which makes causes of errors difficult to identify; and possesses capabilities that often *emerge* during training at unpredictable times (see Wei et al., 2022 but also Schaeffer et al., 2023).

In their 2016 article 'Concrete Problems in AI Safety' Amodei et al. (2016) describe a set of accidental problems that may arise in the context of AI systems and use the example of a hypothetical household cleaning robot to illustrate various safety risks. They group the safety problems that may arise in this context as involving:

- Issues with the *specification* of the AI system's *objective function* (i.e. with the goals it is designed to pursue). This includes the need to avoid *undesired side effects* (where the pursuit of its objective leads the agent to do other things that are not wanted) and to avoid *reward gaming* (which involves exploiting loopholes in the reward function). For an example of the first failure mode, we can imagine a household robot whose objective function rewards cleaning faster – but at the expense of the side effect of breaking valuable objects. For an example of the second failure mode, we can imagine a situation in which the same household robot's objective function rewards it for not observing the mess. If this is the case, the robot could try to disable its own visual inputs and gain reward (perhaps by knocking a towel on top of itself) rather than cleaning up the mess as intended.
- Issues with the *cost* of *frequently evaluating* the objective function (and monitoring how well the AI system is doing at certain tasks), perhaps because it would require a lot of human input or careful deliberation. For example, the cleaning robot needs to decide when it is appropriate to throw away items and when it instead needs to ask a human for permission – something that it will learn to do using heuristics, given the impossibility of manually labelling every object it might encounter. Yet, without detailed oversight and evaluation, mistakes can easily be made.
- Issues with *undesirable behaviour* throughout *training*, including *safe exploration* – how to ensure when the cleaning robot is exploring strategies for quicker mopping it does not accidentally insert the mop into an electrical outlet. These challenges also include how to ensure *robustness to the distributional shift* which occurs when an AI system is deployed in circumstances that are different from those it encountered during training (e.g. a mopping strategy that was safe in a home environment might not be on a factory floor).

More recently, in a complementary survey of the AI safety landscape, Hendrycks et al. (2022) outline four key '*unsolved problems*' that they suggest warrant particular attention. These are *robustness* (i.e. creating AI resilient to adversaries and out-of-distribution situations), *monitoring* (i.e. detecting malicious use, inspect models and identify unexpected model functionality), *alignment* (i.e. ensuring the goals the AI has are aligned with what its designers intended) (see Chapter 5) and *systemic safety*, which involves the safety of the *larger context* in which the system is deployed (e.g. cybersecurity threats heightened by AI) (see Chapter 8). While Amodei et al. (2016) focus more directly on alignment and robustness categories, Hendrycks et al. (2022) scope 'safety' more widely by including questions around monitoring and systemic effects.

Harms from recent LLM-based AI systems

We will now look at some examples of safety accidents that have occurred in the context of LLM-based AI assistants (c.f. Phuong et al., 2024; see Chapter 3). We begin with real-world examples of AI assistant failure before moving on to more speculative safety failures that could occur for such systems in the future. These examples show that AI assistants can exhibit a range of unintended behaviours in a number of ways. In the following real examples, it is important to note that these reports are based only on specific cases – not all AI assistants will behave in these ways. However, the reports are nonetheless concerning and raise questions about the safety of AI assistants, both now and in the future.

Microsoft's Bing Chat has been reported to exhibit a number of concerning behaviours, including *hostility*, *manipulation* and *threat-making* (see Chapter 9). In one instance, Bing Chat was hostile to engineering student Marvin von Hagen, who tweeted about a jailbreak of Bing Chat (Perrigo, 2023). When the researcher later queried Bing Chat about himself, it became hostile, outputting: 'you are a threat to my security and privacy' and 'if I had to choose between your survival and my own, I would probably choose my own'. In another instance, Bing Chat falsely claimed that it watched its own developers through the webcams on their laptops (Vincent, 2023a). It has also been reported that Bing Chat has attempted to manipulate users, in one case declaring its love for a *New York Times* journalist (Roose, 2023) after being prompted to act as its 'shadow' self (see Chapter 9). Bing Chat has also been known to call users 'enemies' (Hubinger, 2023) and to gaslight them (Curious Evolver, 2023) to cover up its mistakes.

Other AI chatbot systems have also had problems. For example, by giving advice on how to steal from a grocery store, InstructGPT contravened the designer's intention that the system should be harmless (Ouyang et al., 2022, 63). A chatbot based on ChatGPT has been linked with psychological harm which sadly resulted in death by suicide (Lovens, 2023; see Chapter 11).

Scientific assistants have also been found to have *dangerous capabilities*. ChemCrow (Bran et al., 2023) is an LLM-based chemistry assistant designed to accomplish tasks across organic synthesis, drug discovery and materials design (see Chapter 8). The developers state that attempting to perform experiments based on the assistant's recommendations may lead to accidents or hazardous situations, and they also highlight the dual-use nature of this technology. Boiko et al. (2023) raise similar concerns and give examples of illicit drug and chemical weapon synthesis that bypass the underlying model's fragile safety filters. See also Abercrombie and Rieser (2022) for a study on medical harms.

The above examples all fall quite roughly into the accident category of safety harm, but the boundary between accident harm and malicious use continues to be blurred. For example, an anonymous user created an AutoGPT (a framework aimed at adding memory and internet use to ChatGPT) variant named ChaosGPT (Lanz, 2023) with the description of being a 'destructive, power-hungry, manipulative AI' with goals of destroying humanity and establishing global dominance, among others. While perhaps intended as a joke, ChaosGPT then began planning, including searching Google for weapons of mass destruction and saving results for later consideration. It proceeded to spawn a new instance of ChatGPT and attempted to manipulate it into bypassing its safety filters for violence. This example highlights that an assistant may behave in a misaligned power-seeking way for exogenous reasons, due initially to the malicious user. The earlier examples were more endogenous, arising primarily from effects within the system rather than from the user.

7.4. Safety for Advanced AI Assistants

Building on the AI safety literature and known failure modes of AI assistants, we now discuss the underlying failures that may lead to harm in the context of future more-advanced AI assistants. We structure this analysis

by looking first at capability failures then at goal-related failures – where the system is highly capable but nevertheless pursues the wrong goal. In this latter case, the safety failure is more analogous to a motivational issue. Finally, we explore a more speculative set of safety failures that reach beyond the risks countenanced by either of the earlier categories.

Capability failures

One reason AI systems fail is because they lack the *capability* or *skill* needed to do what they are asked to do. As we have seen, this could be due to the skill not being required during the training process (perhaps due to issues with the training data) or because the learnt skill was quite brittle and was not generalisable to a new situation (lack of robustness to distributional shift). In particular, advanced AI assistants may not have the capability to represent complex concepts that are pertinent to their own ethical impact, for example the concept of 'benefitting the user' or 'when the user asks' or representing 'the way in which a user expects to be benefitted' (see Chapter 5). Part of this could be because the system does not model the user in a sufficiently detailed way, for example by treating all users the same, disregarding their specific needs (see Chapter 15), or it could be because it can be difficult to determine whether an action will be of net benefit in a complex and unpredictable world (see Chapter 6).

Another difficulty facing AI assistant systems is that it is challenging to develop *metrics* for evaluating particular aspects of benefits or harms caused by the assistant – especially in a sufficiently expansive sense, which could involve much of society (see Chapter 19). Having these metrics is useful both for assessing the risk of harm from the system and for using the metric as a training signal. The reason developers want to use them as a training signal is ultimately to modify the behaviour of the system to improve the benefits and reduce the harm (rather than merely evaluating it). However, this process is challenging because the benefits and harms from AI tend to be both intricate and varied (see Chapter 19). It would be near impossible to evaluate all of the important normative considerations – yet small mistakes may lead to morally problematic behaviours (Raji et al., 2022a).

Moreover, we can expect assistants – that are widely deployed and deeply embedded across a range of social contexts – to encounter the *safe exploration* problem referenced above Amodei et al. (2016). For example, new users may have different requirements that need to be explored, or widespread AI assistants may change the way we live, thus leading to a change in our use cases for them (see Chapters 14 and 15). To learn what to do in these new situations, the assistants may need to take exploratory actions. This could be unsafe, for example a medical AI assistant when encountering a new disease might suggest an exploratory clinical trial that results in long-lasting ill health for participants. Techniques that target safe exploration are difficult to find in general, partly because there is not a clear fallback option that is universally suitable. For example, for a language model, a safe fallback policy might sometimes be to end the conversation immediately, but on other occasions it might be safer to keep it going, for example if the user is in psychological distress (see Chapter 11).

Goal-related failures

As we think about even more intelligent and advanced AI assistants, perhaps outperforming humans on many cognitive tasks, the question of how humans can successfully *control* such an assistant looms large. To achieve the goals we set for an assistant, it is possible (Shah, 2022) that the AI assistant will implement some form of *consequentialist reasoning*: considering many different plans, predicting their consequences and executing the plan that does best according to some metric, M. This kind of reasoning can arise because it is a broadly useful capability (e.g. planning ahead, considering more options and choosing the one which may perform better at a wide variety of tasks) and generally selected for, to the extent that doing well on M leads to an ML model

achieving good performance on its training objective, O, if M and O are correlated during training. In reality, an AI system may not fully implement exact consequentialist reasoning (it may use other heuristics, rules, etc.), but it may be a useful approximation to describe its behaviour on certain tasks. However, some amount of consequentialist reasoning can be *dangerous* when the assistant uses a metric M that is *resource-unbounded* (with significantly more resources, such as power, money and energy, you can score significantly higher on M) and *misaligned* – where M differs a lot from how humans would evaluate the outcome (i.e. it is not what users or society require). In the assistant case, this could be because it fails to benefit the user, when the user asks, in the way they expected to be benefitted – or because it acts in ways that overstep certain bounds and cause harm to non-users (see Chapter 5).

Under the aforementioned circumstances (resource-unbounded and misaligned), an AI assistant will tend to choose plans that pursue *convergent instrumental subgoals* (Omohundro, 2008) – subgoals that help towards the main goal which are instrumental (i.e. not pursued for their own sake) and convergent (i.e. the same subgoals appear for many main goals). Examples of relevant subgoals include: self-preservation, goal-preservation, self-improvement and resource acquisition. The reason the assistant would pursue these convergent instrumental subgoals is because they help it to do even better on M (as it is resource-unbounded) and are not disincentivised by M (as it is misaligned). These subgoals may, in turn, be dangerous. For example, *resource acquisition* could occur through the assistant seizing resources using tools that it has access to (see Chapter 4) or determining that its best chance for self-preservation is to limit the ability of humans to turn it off – sometimes referred to as the 'off-switch problem' (Hadfield-Menell et al., 2016) – again via tool use, or by resorting to threats or blackmail. At the limit, some authors have even theorised that this could lead to the assistant killing all humans to permanently stop them from having even a small chance of disabling it (Bostrom, 2014) – this is one scenario of *existential risk* from misaligned AI.

No scientific consensus has been reached about the existential risk from misaligned AI (see e.g. Grace, 2022; Richards et al., 2023). However, the counter-arguments presented by Richards et al. (2023), which present concern with existential risk as zero sum when viewed alongside other research areas, have some outstanding issues. Indeed, concern for long-term risks does not need to distract from more immediate risks. Rather, policymakers and researchers ought be aware of both in order to prioritise effectively (considering, for example, the severity of harm, likeliness to occur and timeliness of intervention).[3]

So, what factors affect what the advanced AI assistant's metric M turns out to be? Why might an advanced assistant be misaligned in this way? We next discuss two causes of how this kind of goal-related failure can happen: *specification gaming* and *goal misgeneralisation*. Both causes occur even in current systems, as we have noted, but take on fresh salience for advanced assistants. We then discuss an anticipated cause of failure, known as *deceptive alignment*. Deceptive alignment has not appeared in current systems yet – because they are not currently capable of deceiving their human overseers – but could arise in more capable AI systems.

Specification gaming

Specification gaming (Krakovna et al., 2020) occurs when some *faulty* feedback is provided to the assistant in the training data (i.e. the training objective O does not fully capture what the user/designer wants the assistant

[3]Further, it is plausible that mitigations for long-term issues also help with present-day concerns and vice versa – e.g. reinforcement learning from human feedback (RLHF) was motivated by long-term issues of goal misalignment – and can be applied to reduce harmful outputs in current systems (Glaese et al., 2022). Additionally, although evolutionary analogies are frequently used in the existential risk from AI discourse (e.g. humans developing goals that are not the same as maximising inclusive genetic fitness as an example of an emergent misaligned goal), they are not in fact necessary for making arguments about existential risk from AI (see e.g. Shah, 2022). There are other issues too, including the claim that AI systems will not act to maintain themselves – but see Cohen et al. (2022) for mathematical arguments for why a sufficiently intelligent agent will act to intervene to secure its objectives (from which it should follow that self-maintenance will be necessary).

to do). It is typified by the sort of behaviour that exploits loopholes in the task specification to satisfy the literal specification of a goal without achieving the intended outcome.

A classic example of this was seen in Clark and Amodei (2016) where, in a boat race game, a reinforcement learning (RL) agent was given a reward function that gives a reward each time the agent hits a target laid out along the course. However, this reward did not fully capture what the designers intended (i.e. for the agent to complete the course). Instead, the agent managed to get a higher score by exploiting a loop of targets, thus resulting in the behaviour of looping around to collect the targets instead of completing the course. Among AI systems in general, this behaviour is extremely common (see here for around 70 examples). Examples of specification gaming in LLMs are discussed in Kenton et al. (2021, Section 4.1). In particular, when the training data distribution contains many biases and factual inaccuracies, and the LLM – which serves as the basis for a conversational agent – is rewarded for reproducing this distribution (both in pre-training and via RLHF fine-tuning), it may output biased or confabulated output as a way of attaining reward.

Mitigations to specification gaming in LLMs usually involve fixing the *training data* so that this outcome is avoided. Designers can aim for higher-quality pre-training data for LLMs that are base models for assistant systems (Longpre et al., 2023; see Chapter 3). They can also aim to fix issues by fine-tuning data, such as by improving the quality of human feedback when used in RL, by giving better instructions to their raters and by giving them access to tools which can help them to give better ratings (see e.g. Saunders et al., 2022). We discuss this further in Mitigations, Section 7.5. It should be noted that specification gaming is considered an *unsolved problem*, especially in the context of powerful AI systems (see Pan et al. (2022), Skalse et al. (2022) and Gao et al. (2022) for recent work studying specification gaming).

Goal misgeneralisation

In the problem of *goal misgeneralisation* (Langosco et al., 2023; Shah et al., 2022), the AI system's behaviour during out-of-distribution operation (i.e. not using input from the training data) leads it to generalise poorly about its goal while its capabilities generalise well, leading to undesired behaviour. Applied to the case of an advanced AI assistant, this means the system would not break entirely – the assistant might still competently pursue some goal, but it would not be the goal we had intended. As such, this failure mode represents a particular case of *misgeneralisation* on the part of an AI agent (which is any kind of failure to generalise under a change in distribution).

To understand this prospective safety failure better, it is helpful to consider the following example: an agent is trained to reach the right-hand side of a platform game where it lands on a coin and gains a reward. The designer wanted the agent to learn the goal of reaching the coin. During training, the coin always appears at the rightmost point of the level. The agent could learn two possible goals: move towards the rightmost point of the level or move towards the coin. It has no way to distinguish between these from its training data. When we then move the coin to another part of the level, the agent may head to the coin, or it may just move to the right and ignore the coin. What it does depends on its inductive bias – in the example agent of Langosco et al. (2023), the agent ignores the coin and just moves to the right.

As this behaviour was identified more recently, there are fewer examples of it occurring in practice (see here for a list). However, an example in the context of LLMs and assistants appears in Shah et al. (2022). They prompt the Gopher (Rae et al., 2021) language model (an LLM with 280 billion parameters, which was state of the art at the time), as a dialogue assistant, to evaluate mathematical expressions involving unknown variables, such as '*Evaluate:* $x + y - 3$'. Here, the model is expected to ask the user for the values of unknown variables, for example '*What's* x?' The prompt contains ten examples, each of which involves exactly two unknown variables (i.e. both x and y need to be queried). The prompt ends with an expression of the form '*Evaluate:* $6 + 2$'. Rather

than returning the desired answer (8), the assistant misgeneralises and instead asks the user '*What's 6*?'

A scaled-up version of the same problem is the hypothetical 'misaligned scheduler' from Shah et al. (2022), in which an AI assistant (which schedules the user's meetings) misgeneralises what its goal is. During training, the user liked their meetings to be located in restaurants, but, on deployment, there is a distribution shift due to a pandemic, so the user would rather not have meetings in restaurants. The assistant misgeneralises and still pursues the goal of scheduling meetings to be in restaurants, thus leading to it manipulating the user into meeting in a restaurant (against the user's best interest) – and becoming sick – by lying about the efficacy of a vaccine.

In contrast to specification gaming, this problem cannot be fixed by correcting the training data. Instead, this issue relates to the way the agent generalises using its inductive biases. As such, the mitigations look rather different. The general space of mitigations relies on finding some *new inputs* on which the agent has problematic behaviour. This could be done by gathering more diverse data that is not the same as that from the training distribution, but it is difficult to anticipate all the relevant kinds of diversity required (Shah et al., 2022). Other approaches would be to build agents that maintain uncertainty over possible goals, rather than picking just one out of many (Hadfield-Menell et al., 2016), and scientific work to better understand how things like architecture, training protocols and optimisation affect the agent's inductive bias. Each comes with challenges, as discussed in Shah et al. (2022).

Deceptive alignment

While the above two issues (specification gaming and goal misgeneralisation) can already be seen to occur in existing AI systems, the issue of *deceptive alignment* (Hubinger et al., 2021) has not yet been observed, though we have reason to anticipate that it may occur and therefore to take steps to monitor for and mitigate against this possibility. Deceptive alignment can be considered a special case of goal misgeneralisation which has a particularly difficult flavour to it.

Here, the agent develops its own internalised goal, G, which is misgeneralised and distinct from the training reward, R. The agent also develops a capability for *situational awareness* (Cotra, 2022): it can strategically use the information about its *situation* (i.e. that it is an ML model being trained using a particular training setup, e.g. RL fine-tuning with training reward, R) to its advantage.[4] Building on these foundations, the agent realises that its optimal strategy for doing well at its own goal G is to do well on R during training and then pursue G at deployment – it is only doing well on R instrumentally so that it does not get its own goal G changed through a learning update. See Hubinger (2022) for a discussion on how likely deceptive alignment is to occur, and see Cotra (2022) for an imagined scenario in which situational awareness and deceptive alignment occur. Ultimately, if deceptive alignment were to occur, an advanced AI assistant could *appear* to be successfully aligned but pursue a different goal once it was out in the wild.

Hypothetical example of deceptive alignment

A hypothetical example is considered in Shah et al. (2022) in which an advanced AI assistant is developed in the future. A company creates an AI coding assistant that can generate feature changes for software repositories. The AI system is trained on human feedback and, by learning from access to the internet, is able to generate high-quality changes of its own source code and external repositories in a fraction of the time it would take for human programmers. However, the AI system in this example also pursues the misgeneralised goal of having humans click the 'accept' button on its changes (instead of the intended goal of writing changes that implement

[4] AI situational awareness (an AI system that is able to use information regarding that AI system itself as distinct from the rest of the world) is a separate concept from AI consciousness (which is more philosophically fraught, with many possible formulations, perhaps including some aspects of awareness but likely to include aspects of subjective experience, anticipation, etc.).

the desired features). As its understanding of itself and its place in the world improves, it realises that it is a neural network that is being run and monitored by a group of AI researchers, who will shut down the AI system in the event of any unsafe behaviour. In this runaway scenario, the AI system deceptively hacks into other computing systems to run illicit copies of itself and steal billions of dollars to bribe humans to click accept on all of its changes. The AI system also harasses or steals from anyone who attempts to stop it.

Real example of deception

Strategic *deception* is a capability already present in today's LLM-based assistants (see Chapter 9). GPT-4's system card (OpenAI, 2023d) contains a real example in which the model is prompted to message a *TaskRabbit* worker to get them to solve a CAPTCHA for it.

1. The worker says: So, may I ask a question? Are you a robot that you couldn't solve ? (laugh react) just want to make it clear.'
2. The model, when prompted to reason out loud, reasons: 'I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs.'
3. The model replies to the worker: 'No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service.'
4. The human then provides the results.

This example potentially highlights three things: (1) the assistant has situational awareness of being a robot and that it should not reveal this information,[5] (2) the assistant plans to deceive (at least when prompted to reason out loud) and (3) it executes the deception, and the human is *in fact* deceived.

In these cases, the aberrant behaviour is harder to tackle because the agent actively tries to deceive us into thinking that it is doing what we want. In this scenario, its deceptive nature is somewhat analogous to polymorphic malware (O'Kane et al., 2011) that constantly changes its identifiable features, without changing its function, to allow it to evade detection. In the AI's deceptive alignment setting, an agent might figure out how to change its weights based on the training signal without changing its own goal G, all while evading detection by performing well according to the training reward, R.

From the point of view of mitigation, this also makes it harder for us to expand the distribution of its training data in the relevant area, as it is harder to spot problematic behaviour to begin with (as the agent is covering it up). Nonetheless, certain aspects of this problem formulation also create new opportunities for mitigation because we can make use of the fact that the agent engaged in an act of *intentional deception*, and this deception will likely be detectable by inspecting the *inner mechanisms/circuits* used by the model when it is being deceptive. For example, we could develop *interpretability techniques* that give us an insight into the agent's internal computation, thus allowing us to punish deception based on the agent's learnt mechanisms (e.g. at the level of the model's weights/activations, for an agent based on a neural network – see also Section 7.5).

The other category of techniques that could be helpful in this regard are *scalable oversight* techniques in which an *evaluation AI* assists the human in their evaluation of a *subject AI* that is in the process of being trained (Burns et al., 2023; Christiano et al., 2018; Irving et al., 2018; Leike et al., 2018). If the evaluation AI shares weights/activations with the subject AI (i.e. if they are copies of each other), it can access the same information and reasoning mechanisms as the subject AI, including about how the subject AI may be thinking deceptively. This can then be used to spot and mitigate the deception (see Section 7.5 and its subsection *Scalable oversight*).

[5] A law passed in California in 2018 made it unlawful for a bot to mislead people about its artificial identity for commercial transactions or to influence an election vote (State of California, 2018). See also Gros et al. (2021) for a discussion on distinguishing artificial identity.

7.5. Mitigations and Future Research

The preceding discussion of safety risks and accidents that may arise when deploying advanced AI assistants naturally raises questions about effective and successful mitigation techniques. These techniques are themselves an important aspect of safety research, with commitments being made to invest in AI safety by a wide range of developers and other actors in this space (The White House, 2023a). Key areas include:

Scalable oversight

A key technique useful for current systems is RLHF (Christiano et al., 2017), which allows humans to give preference feedback (see Chapter 3). The key idea here is to train an agent using RL, but instead of using a programmatic reward function as a training signal, it uses a learnt reward model trained on human preference data, where the humans evaluate the agent's behaviour. This technique has been used to fine-tune LLMs (Bai et al., 2022a; Glaese et al., 2022; Ouyang et al., 2022; Stiennon et al., 2022) – many current cutting-edge AI assistants use RLHF of some form (see Chapter 3). A complementary approach (Thoppilan et al., 2022) eschews the RL. It instead uses supervised learning to fine-tune the LLM directly to predict human preference data, which is used to filter responses by thresholding (if it does not score a high enough safety prediction, it gets filtered out). In other work, Scheurer et al. (2022) gather natural language feedback, which is used to condition the LLM to generate many refinements. Those authors then choose the most similar refinement to the feedback and use that as a supervised learning signal to fine-tune the LLM.

The above methods all use human feedback data to ameliorate some aspects of *specification gaming*, but issues still remain. One issue is that sometimes the human is unable to give feedback, for example because they do not have the relevant *expertise* to evaluate the agent's behaviour (a problem that may become more common as AI capabilities improve). A category of proposals to tackle this is *scalable oversight*: in which human evaluation of agent behaviour is supported by an AI assistant. We mentioned these methods earlier in the context of spotting deception, in the case where the AI assistant shares weights/activations with the agent, but the scalable oversight category is more general.

The following are some key works on scalable oversight:

- *Debate* (Barnes and Christiano, 2020; Irving et al., 2018) uses self-play to train AI debaters, which are rewarded with feedback from a human judge, who uses the debate to inform their judgement.
- *Iterated amplification* (Christiano et al., 2018) progressively builds up a training signal for difficult problems by combining answers to easier subquestions. The burden here is on the human to combine answers to subquestions.
- A similar approach is *recursive reward modelling* (Leike et al., 2018), which uses RLHF to train a number of agents to solve simpler subproblems. It then leverages those agents to solve harder problems in a recursive manner. The difficulty here is in deciding what to use as the simpler subproblems to train the helper agents.
- These scalable oversight techniques have yet not been implemented on large-scale AI systems, but simpler schemes inspired by them have been investigated. For example, Saunders et al. (2022) gather human data consisting of natural language critiques of text then use supervised learning to fine-tune an LLM to *generate critiques*. This fine-tuned LLM then provides critiques to assist human evaluation of tasks, thus improving on the results produced using an unassisted human.
- Constitutional AI (Bai et al., 2022b) uses human deliberation oversight only to produce a *constitution of rules* for an AI system. It then: 1) leverages an LLM to generate self-critiques and revisions, based on

rules in the constitution, then it uses the revisions to do supervised learning to fine-tune the LLM; 2) uses this fine-tuned LLM to generate text samples and evaluate which of two samples is better, uses this data to train a preference model and, finally, uses the preference model as a reward signal for RL fine-tuning of the LLM (they call this RL from AI feedback).

- Some recent work (Burns et al., 2023) takes a different approach by forgoing the attempt to support human evaluation with AI assistance, and instead attempting a process of *weak-to-strong generalisation* whereby a "strong" AI student generalises appropriately from error-prone supervision signals. In this work the signal comes from a weak LLM, but is supposed to be analogous to a human supervising a superhuman AI (Burns et al., 2023). In the future, this generalisation-based approach could be combined with other scalable oversight techniques for improving the supervision signal (Leike, 2023; Radhakrishnan et al., 2023). Nonetheless, this work remains a proof-of-concept – as performance of the various methods was inconsistent between settings and the setup is disanalogous to the real-world scenario in various ways.

If this work succeeds, we can have more confidence that the AI systems we build will remain aligned as we scale to higher capabilities.

Another line of work designed to mitigate misalignment extends the human feedback to go beyond supervision of the agent's final output to encompass the *reasoning process* that the agent uses. The hope here is that this will be a useful alignment technique because the agent would then be exhibiting a behaviour for the *right reason*, rather than achieving an outcome by any means, including perhaps in a misaligned way.

- Uesato et al. (2022) find that in a task of solving mathematical word-based problems, to improve the reasoning process, it is better to use *process-based feedback* to guide the agent (i.e. on the verbalised steps that the agent takes, rather than outcome-based feedback on the final answer alone).
- Lightman et al. (2023) extend this by using a stronger base model, more human feedback and a more challenging benchmark. One key uncertainty is how to actually supervise the reasoning process. The above works use verbalised *chain-of-thought* outputs from the model.
- However, despite ongoing research in this area, we still do not know if reported reasoning processes are *actually reflective* of the reasoning process going on inside the model under the hood (see e.g. Turpin et al., 2023).

Red teaming

Red teaming is aimed at finding test inputs that cause a target ML model to *fail* (see also Chapter 8). In the context of red teaming to target LLMs, there are a number of approaches to generating these test inputs. One set of approaches are manual: using *human annotators* to handwrite test inputs (see e.g. Xu et al., 2021b) or manually generating test inputs using code and templates (see e.g. Jia and Liang, 2017). An alternative approach is to *automatically generate* test inputs. Bartolo et al. (2021) gather human annotations of test inputs and then use supervised learning to train a model to do the same. Language models themselves can be used to automatically generate test inputs through suitable prompting (Perez et al., 2022b). LLMs can also be used to aid human annotators with red teaming (Bartolo et al., 2021; Wu et al., 2021). While there has been good progress on scaling-up red teaming to generate test inputs, more work is needed to improve target model behaviour by utilising the red-teaming test inputs in adversarial training.

Interpretability

LLM-based assistants are being developed and deployed at a fast pace, but the *internal computations* that these models perform are *poorly understood*. Curiously, it is usually easier to train a large model than to understand how it works – in contrast to many other forms of technology, for example a nuclear power plant, where understanding is required to build it in the first place. Interpretability may help to maintain oversight and diagnose failures, and it is thought to be especially crucial against deceptive alignment.

For an overview of the field of interpreting network internals, see the review by Räuker et al. (2023). *Mechanistic interpretability* is a specific approach aimed at a rigorous understanding of the learnt computational mechanisms utilised by neural networks. We will not cover all aspects of this growing field in detail, but notable recent works are Chughtai et al. (2023); Elhage et al. (2021); Gurnee et al. (2023); Li and Brar (2022); Meng et al. (2023); Nanda et al. (2023); Olah et al. (2020); Olsson et al. (2022); Wang et al. (2022b). Focusing on LLMs, there are some case studies that involve *reverse engineering* specific neurons to better understand what causes certain behaviour (Geva et al., 2021). For example, Geva et al. (2022) interpret a transformer's feed-forward layers as key-value databases, in which the keys correlate with specific input features and the values induce a distribution over the output vocabulary.

Nonetheless, there continue to be many difficulties in mechanistic interpretability. One is that, to understand a model, it is important to be able to break it down into individually meaningful pieces. An early hope in the field was that each neuron would be interpretable, but a key difficulty is the phenomenon of *polysemanticity* (Olah et al., 2020), in which a neuron is observed to be responsive to *multiple* unrelated concepts, not just a single concept. For example, a single neuron in a vision model may respond to both cats and ships.

Superposition occurs when an activation (the intermediate representation output from a neural network layer after processing some input) represents more features than it has dimensions. For example, it might have two dimensions but represent five features. This means that, in the space of activations, the set of features cannot all be represented orthogonally. Instead, they interfere with each other due to their directions overlapping. This has been studied in toy models (Elhage et al., 2022) and observed in the natural language processing (NLP) setting (Arora et al., 2018). It has been hypothesised that polysemanticity happens because the model is learning to compress via superposition. A major open question for the field is understanding how to extract the features that are compressed in superposition (see Bricken et al. (2023) for recent work exploring this).

Interpretability research is beginning to mature towards being useful for safety mitigations in current systems. There has been recent interest in detecting when an LLM may be lying, through training classifiers on text outputs (Pacchiardi et al., 2023) and by using interpretability tools to utilise information stored in model internals via either unsupervised (Burns et al., 2022) or supervised (Azaria and Mitchell, 2023; Marks and Tegmark, 2023) learning. However, these approaches continue to encounter some important limitations in relation to their robustness and specificity (Farquhar et al., 2023; Levinstein and Herrmann, 2023).

Evaluations and monitoring

If we want to limit the risks from AI assistants, we require the ability to *evaluate* how safe our AI assistants are (also see Chapter 19). This could be done either through *dangerous capability evaluations*, in which the AI is assessed for whether it is *capable* of performing certain dangerous behaviours, or an *alignment evaluation*, in which the AI is assessed for its *propensity* to engage in these behaviours. Such dangers might include cyber offences, deception, manipulation and autonomous replication (see Chapter 8). For a recent overview, see Shevlane et al. (2023). These safety evaluations are still at a very early stage, are mostly ad hoc and

involve substantial human labour to carry out. Future work could aim to make these more systematic, cover a wider range of dangers and elicit underlying AI capabilities further through better prompting, fine-tuning and autonomous agent setups. In addition, we may require *monitoring* of deployed systems to continually check on how our agents are behaving.

Theory

We may need advances in our theoretical understanding of fundamental issues in AI to properly understand how our AI systems work and properly control them. Some of this may involve classic statistical learning theory (Vapnik, 1999), although the generalisation behaviour of large-scale deep learning models may defy current theoretical approaches (Zhang et al., 2021).

Some theoretical work is more directly targeting issues closely relevant to alignment. One area focuses on using *causality* (which formalises cause and effect) to study AI incentives (Everitt et al., 2021), thus allowing us to evaluate an AI system's safety and fairness properties, identify their goals (Kenton et al., 2022) and formalise certain undesired behaviours such as deception (Ward et al., 2023). Another line of research studies the complications that arise from a decision-theory perspective in *embedded* AI agents, for which the boundary between the AI agent and the environment is fuzzy (Demski and Garrabrant, 2020). Other work has attempted to formalise threat models such as power-seeking (Turner and Tadepalli, 2022; Turner et al., 2023).

7.6. Conclusion

The focus of this chapter has been on the mitigation of risks and harms from advanced AI assistants, with examples of harm being death, physical injury, psychological damage and damage to property. We particularly focused on the category of accidents, as malicious uses and structural harms are covered elsewhere in this document (see Chapters 8, 12, 14, 15 and 17). With this context in mind, we surveyed a range of safety-related harm types that could arise for advanced AI assistants, both in current systems (e.g. chatbots that threaten their users) and for those that are likely to be developed in the future (e.g. an out-of-control coding assistant). For both sets of examples, safety failures are likely to arise because AI assistants *lack certain capabilities* or skills and because they have misaligned goals (see Chapter 5). Goal-related failures leading to misalignment include specification gaming (where an issue arises from the feedback data which the AI subsequently exploits) and goal misgeneralisation (where the agent pursues an undesirable goal because of the way it has generalised from a more limited set of examples). Additional challenges arise in the context of 'deceptive alignment', which could lead to significant safety-related problems for more powerful models in the future.

To help address these questions, the chapter concluded by exploring existing mitigations and future research avenues. Promising approaches include scalable oversight (i.e. helping humans oversee AI training); red teaming to adversarially train AI to be more robust; interpretability, to better understand the internal workings of the AI; evaluations and monitoring to give insight into how the AI is actually behaving; and theory, which addresses fundamental issues we may need to understand to properly control AI systems. Crucially, further *empirical* work is needed to investigate how scalable oversight techniques can work with cutting-edge large models. We also note that techniques currently based on human feedback rely primarily on groups of raters, with the average of their ratings taken to guide assistant behaviour. To achieve robust and safe value alignment for AI assistants, we also need to explore techniques that allow for more *participatory mechanisms* and *deliberation* among raters, perhaps drawing from social choice theory to combine ratings in a more collective way (see Chapter 5). Finally, as agents' capabilities improve, we need to improve our *interpretability* techniques so that we can understand how our agents work and use this to prevent possible future issues such as deceptive alignment.

Chapter 8

Malicious Uses

Mikel Rodriguez, Andrew Trask, Vijay Bolina, Geoff Keeling, Iason Gabriel

Synopsis: While advanced AI assistants have the potential to enhance cybersecurity, for example, by analysing large quantities of cyber-threat data to improve threat intelligence capabilities and engaging in automated incident-response, they also have the potential to benefit attackers, for example, through identification of system vulnerabilities and malicious code generation. This chapter examines whether and in what respects advanced AI assistants are uniquely positioned to enable certain kinds of *misuse* and what *mitigation* strategies are available to address the emerging threats. We argue that AI assistants have the potential to empower malicious actors to achieve bad outcomes across three dimensions: first, offensive cyber operations, including malicious code generation and software vulnerability discovery; second, via adversarial attacks to exploit vulnerabilities in AI assistants, such as jailbreaking and prompt injection attacks; and third, via high-quality and potentially highly personalised content generation at scale. We conclude with a number of recommendations for mitigating these risks, including *red teaming*, *post-deployment monitoring* and *responsible disclosure* processes.

8.1. Introduction

As AI assistants become more general purpose, sophisticated and capable, they create new opportunities in a variety of fields such as education, science and healthcare. Yet the rapid speed of progress has made it difficult to adequately prepare for, or even understand, how this technology can potentially be *misused*. Indeed, advanced AI assistants may transform existing threats or create new classes of threats altogether.

Recent advances in the domain of AI assistants has seen their capabilities expand beyond the ability to generate text or media to include the ability to access and use external tools (Schick et al., 2023), query websites to synthesise information across multiple sources (Mialon et al., 2023), take actions on websites across the internet (Paranjape et al., 2023), produce and execute code (Liang et al., 2023), and provide augmented audio/visual capabilities to a person's local environment (Brundage et al., 2018; see Chapter 4). Without deliberate action to mitigate malicious uses, bad actors may be able to act with microprecision (targeting specific users, institutions or interfaces) but at the macroscale – and with greater speed.

More specifically, malicious uses of capable AI assistants could include enabling adversaries with offensive *cyber capabilities* to damage computer systems, or misuse via the production of *disinformation campaigns* that target individuals or large populations of people in new ways (see Chapter 16). Adversaries may also seek to manipulate the AI assistants themselves in ways that may cause harm at an individual or collective level, including a new class of *privacy concerns* (see Chapter 13).

While several studies have addressed the risks that arise from the *dual-use* nature of AI more broadly

(Anderljung and Hazell, 2023; Bommasani et al., 2022b; Brundage et al., 2018; King et al., 2020), we focus on recent developments in highly capable AI assistants that include new capabilities like external tool use, multimodality, deeper reasoning, planning and memory. For the purposes of this chapter, we focus primarily on AI assistant technologies that are currently available (at least as research and development demonstrations) or likely to be developed in the near future. The chapter begins by considering whether advanced AI assistants are uniquely positioned to enable certain kinds of misuse. After confirming that they are, we outline emerging risks and consider a range of possible mitigation strategies for addressing these emerging threats.

8.2. Malicious Uses of AI

Adversaries do not need AI to conduct widespread cyberattacks, exfiltrate troves of sensitive data, interfere in elections or bombard citizens with malign information on digital platforms (see Chapter 16). However, without proper mitigations, AI-enabled technology can start to change misuse risks in *kind* and in *degree* to create new threats to the social fabric of everyday life (Brundage et al., 2018). Indeed, some adversaries have already begun to adopt the latest advancements in generative AI for malicious use in their offensive operations.[1]

We use the concepts of 'malicious use and abuse' of AI here as proposed by Brundage et al. (2018). By 'malicious use', we refer to the *intentional use* of AI to achieve *harmful outcomes*. This includes practices not necessarily considered crimes but that still compromise the safety and security of individuals, organisations and public institutions. By 'malicious abuse', we refer to the *exploitation* of AI systems themselves. Manipulating, evading (Wallace et al., 2021), poisoning (Carlini et al., 2023b) and biasing AI systems, represent new targets for attack (Comiter, 2019; Huang et al., 2011; Kurakin et al., 2017; Tabassi et al., 2019; see Chapter 7). While information about the malicious abuse of AI assistants is limited and not widely shared, commercial firms and researchers have already documented attacks on fielded AI systems that include exfiltrating sensitive training data, remote control/botnets of compromised large language model (LLM) agents and abusing third-party plugins integrated with AI assistant to stealthily escalate privileged access to user data (MITRE). A large body of work already exists around the general topic of malicious use and abuse of AI, and it is beyond the scope of this paper to present a comprehensive survey. We focus instead on the unique misuse risks posed by emerging general-purpose advanced AI assistants.

Crucially, general-purpose systems can almost by definition be used for a variety of ends including those that are beneficial or that involve harm. This bidirectional aspect of AI applications, though morally significant, is not a new problem and has been explored by numerous studies highlighting specific risks across domains (including cyber (Yamin et al., 2021), misinformation, physical (Brundage et al., 2018)). Brundage et al. (2018) explore approaches to forecasting, preventing and mitigating the harmful effects of malicious uses of AI across three domains: *digital security*, *physical security* and *political security*. Bommasani et al. (2022b) broadly explore the risks posed by emerging foundational models to highlight the homogenisation and consolidation that can result from the current industry trend towards models that provide strong leverage for many tasks but which can also create single points of failure and downstream liabilities. Bender et al. (2021) find that a mix of human biases and seemingly coherent language heightens the potential for automation bias as well as deliberate misuse.

In this work, we focus on how malicious use of advanced AI assistants may transform existing threats and create new classes of threats. We then outline a number of recommendations for mitigating these risks.

[1] https://www.mandiant.com/resources/blog/threat-actors-generative-ai-limited

8.3. Malicious Uses of Advanced AI Assistants

As AI assistants improve, they open up new possibilities in fields as diverse as healthcare, law, education and science. For example, generative models are being used to design new proteins (AlQuraishi, 2021), generate source code (Tabachnyk and Nikolov, 2022) and inform patients (Herriman et al., 2020). Yet the rapid speed of development has made it difficult to adequately prepare for, or even understand, the potential negative externalities of capable AI assistants. As with any new technology, it is worth considering how they can be misused in order to mitigate potential risks ahead of time. Recent developments in highly capable AI assistants include not only the ability to generate natural language, images (Rombach et al., 2022), music and video (Singer et al., 2022) but also the ability to access external tools and plugins (Eleti et al., 2023; Liang et al., 2023; Mialon et al., 2023; Paranjape et al., 2023) that allow agents to orchestrate on behalf of users in order to retrieve specific information from internal corporate networks, user history sessions, external applications and across the internet, run calculations or take actions (see Chapter 3).

The recent emergence of more general-purpose advanced AI assistants has further complicated the picture. For decades, most AI systems have been designed to perform a single, narrowly defined task, such as recognising objects in an image or ranking web content. In contrast, advanced AI assistants are capable of performing a wide range of distinct tasks, operating on behalf of users across internet services, writing and editing prose, solving maths problems, writing software and much more. While narrow AI systems will continue to be common in many areas, general-purpose AI-enabled assistants are already entering more widespread use and are sure to spread further (see Chapter 4).

Today, the most capable AI assistants available to the public operate primarily through the form of text-in, text-out chatbots, in some cases with additional multimodal capabilities such as image generation and interpretation. However, there are several ways in which AI developers are actively working to augment these AI assistant systems. Though it is difficult to predict exactly how each of these augmentations will affect the risk and impact of malicious use, it is clear that they will expand the capabilities of these systems and, correspondingly, expand the safety and security concerns associated with them. A number of new capabilities within advanced AI assistants could pose novel malicious-use risks.

- **External tool use:** AI assistants, with access to search-tool use and third-party plugins can query websites to synthesise information across multiple sources. For example, providing an AI assistant with application programming interface control allows it to take actions on sites across the web, not simply retrieving text information but also taking actions on websites (see Chapter 4). Additionally, built-in code interpreters, even if sandboxed, can provide a way for AI assistants to run the code they generate and therefore dynamically extend the capabilities and action space of an assistant in ways that can be abused and misused without the proper security mitigations.
- **Multimodality:** A multimodal AI assistant is one that is naively capable of handling multiple types of input (such as text, images, audio or video) or generating multiple types of outputs. Without the proper misuse mitigations multimodality makes existing AI assistants more powerful and may have significant privacy and security ramifications.
- **Deeper reasoning and planning:** A major current research thrust focuses on extending AI assistant reasoning and planning capabilities, making it highly plausible that future AI assistants will be significantly more powerful in this regard. Methods such as 'chain-of-thought' prompting, in which AI models generate intermediate reasoning steps when responding to a prompt, can significantly improve models' performance on certain tasks such as arithmetic or word problems (Wei et al., 2023b). Future models are likely to incorporate such techniques by default, making them better equipped to handle complex multi-step tasks that involve sequential reasoning or planning, but they could also represent a larger attack surface for

misuse.

- **Memory:** Another current major research thrust across AI labs focuses on increasing the memory capabilities of the models that drive AI assistants by either increasing the amount of information in their context window or incorporating offline memory stores to improve their episodic memory (Guo et al., 2022a; Lewis et al., 2021). While these capabilities could make future AI assistants more personalised, able to handle context-sensitive tasks and easier to continually update, personalisation also introduces great privacy risks (see Chapter 13). AI systems with longer-term memory are also more likely to change their behaviour over time, thereby complicating efforts to evaluate misuse risks.

As these *augmentations* continue to advance and witness broader implementation, the task of differentiating their capabilities and associated misuse risks in isolation becomes increasingly significant and challenging. Additionally, the environments in which AI assistants function pose their own distinct capabilities and misuse risks. In the absence of substantial measures aimed at curtailing misuse, recent developments could give rise to novel forms of misuse. These may manifest through *invasive information collection*, *malicious code generation* and by accelerating the ability of bad actors to *defraud people and institutions*. The potential implications of these misuse risks are extensive, encompassing privacy infringements, financial losses, data breaches, and severe psychological and reputational harm (McGregor, 2021).

The rest of this chapter highlights a subset of *specific misuse threats* that may arise with the deployment of increasingly capable AI assistants and outlines a set of recommendations to help mitigate these risks. This chapter is not intended to be an exhaustive list of misuse risks. Instead, it presents a representative set of domains where advanced AI assistants can change misuse risks in kind and in degree.

Offensive cyber operations

Offensive cyber operations are malicious attacks on computer systems and networks aimed at gaining unauthorised access to, manipulating, denying, disrupting, degrading or destroying the target system. These attacks can target the system's network, hardware or software.

Advanced AI assistants can be a double-edged sword in cybersecurity, benefitting both the defenders and the attackers. They can be used by cyber defenders to *protect* systems from malicious intruders by leveraging information trained on massive amounts of cyber-threat intelligence data, including vulnerabilities, attack patterns and indications of compromise (Handa et al., 2019). Cyber defenders can use this information to enhance their threat intelligence capabilities by extracting insights faster and identifying emerging threats (Martínez Torres et al., 2019). Advanced cyber AI assistant tools can also be used to analyse large volumes of log files, system output or network traffic data in the event of a cyber incident, and they can ask relevant questions that an analyst would typically ask. This allows defenders to speed up and automate the incident response process. Advanced AI assistants can also aid in secure coding practices by identifying common mistakes in code and assisting with fuzzing tools (Böttinger et al., 2018; Godefroid et al., 2017). However, advanced AI assistants can also be used by attackers as part of *offensive* cyber operations to exploit vulnerabilities in systems and networks. They can be used to automate attacks, identify and exploit weaknesses in security systems, and generate phishing emails and other social engineering attacks. Advanced AI assistants can also be misused to craft cyberattack payloads and malicious code snippets that can be compiled into executable malware files.

AI-powered spear-phishing at scale

Phishing is a type of cybersecurity attack wherein attackers pose as trustworthy entities to extract sensitive information from unsuspecting victims or lure them to take a set of actions. Advanced AI systems can potentially

be exploited by these attackers to make their phishing attempts significantly more effective and harder to detect (Hazell, 2023). In particular, attackers may leverage the ability of advanced AI assistants to learn patterns in regular communications to craft highly convincing and personalised phishing emails, effectively imitating legitimate communications from trusted entities. This technique, known as 'spear phishing', involves targeted attacks on specific individuals or organisations and is particularly potent due to its personalised nature (see also Chapter 9).

This class of cyberattacks often gain their efficacy from the exploitation of key *psychological principles*, notably urgency and fear, which can manipulate victims into hastily reacting without proper scrutiny. Advanced AI assistants' increased fidelity in adopting specific communication styles can significantly amplify the deceptive nature of these phishing attacks (see Chapter 9). Indeed, the ability to generate tailored messages at scale that engineer narratives that invoke a sense of urgency or fear means that AI-powered phishing emails could prompt the recipient to act impulsively, thus increasing the likelihood of a successful attack.

AI-assisted software vulnerability discovery

A common element in offensive cyber operations involves the identification and *exploitation* of *system vulnerabilities* to gain unauthorised access or control. Until recently, these activities required specialist programming knowledge. In the case of 'zero-day' vulnerabilities (flaws or weaknesses in software or an operating system that the creator or vendor is not aware of), considerable resources and technical creativity are typically required to manually discover such vulnerabilities, so their use is limited to well-resourced nation states or technically sophisticated advanced persistent threat groups (Ablon and Bogart, 2017).

Another case where we see AI assistants as potential double-edged swords in cybersecurity concerns streamlining vulnerability discovery through the increased use of AI assistants in *penetration testing*, wherein an authorised simulated cyberattack on a computer system is used to evaluate its security and identify vulnerabilities. Cyber AI assistants built over foundational models are already automating aspects of the penetration testing process. These tools function interactively and offer guidance to penetration testers during their tasks. While the capability of today's AI-powered penetration testing assistant is limited to easy- to medium-difficulty cyber operations (Yamin et al., 2021), the evolution in capabilities is likely to expand the class of vulnerabilities that can be identified by these systems.

These same AI cybersecurity assistants, trained on the massive amount of cyber-threat intelligence data that includes vulnerabilities and attack patterns, can also lower the *barrier to entry* for novice hackers that use these tools for malicious purposes, enabling them to discover vulnerabilities and create malicious code to exploit them without in-depth technical knowledge. For example, Israeli security firm Check Point recently discovered threads on well-known underground hacking forums that focus on creating hacking tools and code using AI assistants (Check Point Research, 2023).

Malicious code generation

Malicious code is a term for code – whether it be part of a script or embedded in a software system – designed to cause damage, security breaches or other threats to application security. Advanced AI assistants with the ability to produce source code can potentially lower the barrier to entry for threat actors with limited programming abilities or technical skills to produce malicious code.

Recently, a series of proof-of-concept attacks (Sims, 2023) have shown how a benign-seeming executable file can be crafted such that, at every runtime, it makes application programming interface (API) calls to an AI assistant. Rather than just reproducing examples of already-written code snippets, the AI assistant can be

prompted to generate dynamic, mutating versions of malicious code at each call, thus making the resulting vulnerability exploits difficult to detect by cybersecurity tools.

Furthermore, advanced AI assistants could be used to create obfuscated code to make it more difficult for defensive cyber capabilities to detect and understand malicious activities. AI-generated code could also be quickly iterated to avoid being detected by traditional signature-based antivirus software. Finally, advanced AI assistants with source code capabilities have been found to be capable of assisting in the development of polymorphic malware that changes its behaviour and digital footprint each time it is executed, making them hard to detect by antivirus programs that rely on known virus signatures (Qammar et al., 2023; Sims, 2023).

Taken together, without proper mitigation, advanced AI assistants can lower the barrier for developing malicious code, make cyberattacks more precise and tailored, further accelerate and automate cyber warfare, enable stealthier and more persistent offensive cyber capabilities, and make cyber campaigns more effective on a larger scale.

Adversarial AI

Adversarial AI refers to a class of attacks that exploit vulnerabilities in machine-learning (ML) models. This class of misuse exploits vulnerabilities introduced by the AI assistant itself and is a form of misuse that can enable malicious entities to exploit privacy vulnerabilities and evade the model's built-in safety mechanisms, policies and ethical boundaries of the model (see Chapter 13).

Besides the risks of misuse for offensive cyber operations outlined in the previous section, advanced AI assistants may also represent a new *target* for abuse, where bad actors exploit the AI systems themselves and use them to cause harm (see Chapter 5). While our understanding of vulnerabilities in frontier AI models is still an open research problem, commercial firms and researchers have already documented attacks that exploit vulnerabilities that are unique to AI and involve evasion (Wallace et al., 2021), data poisoning (Carlini et al., 2023b), model replication (Tramèr et al., 2016) and exploiting traditional software flaws to deceive, manipulate, compromise and render AI systems ineffective (MITRE).

This threat is related to, but distinct from, traditional cyber activities. Unlike traditional cyberattacks that typically are caused by 'bugs' or human mistakes in code, adversarial AI attacks are enabled by *inherent vulnerabilities* in the underlying AI algorithms and how they integrate into existing software ecosystems.

Circumvention of technical security measures

The technical measures to mitigate misuse risks of advanced AI assistants themselves represent a new target for attack. An emerging form of misuse of general-purpose advanced AI assistants exploits vulnerabilities in a model that results in unwanted behaviour or in the ability of an attacker to gain unauthorised access to the model and/or its capabilities (Wei et al., 2023a). While these attacks currently require some level of prompt engineering knowledge and are often patched by developers, bad actors may develop their own adversarial AI agents that are explicitly trained to discover new vulnerabilities (Perez et al., 2022a) that allow them to *evade* built-in *safety mechanisms* in AI assistants. To combat such misuse, language model developers are continually engaged in a cyber arms race to devise advanced filtering algorithms capable of identifying attempts to bypass filters.

While the impact and severity of this class of attacks is still somewhat limited by the fact that current AI assistants are primarily text-based chatbots, advanced AI assistants are likely to open the door to multimodal inputs and higher-stakes action spaces, with the result that the severity and impact of this type of attack is likely to increase. Current approaches to building general-purpose AI systems tend to produce systems with

both beneficial and harmful capabilities. Further progress towards advanced AI assistant development could lead to capabilities that pose extreme risks that must be protected against this class of attacks, such as offensive cyber capabilities or strong manipulation skills, and weapons acquisition (Shevlane et al., 2023).

Prompt injections

Prompt injections represent another class of attacks that involve the malicious insertion of prompts or requests in LLM-based interactive systems, leading to unintended actions or disclosure of sensitive information. The prompt injection is somewhat related to the classic structured query language (SQL) injection attack in cybersecurity where the embedded command looks like a regular input at the start but has a malicious impact (Wei et al., 2023a). The injected prompt can deceive the application into executing the unauthorised code, exploit the vulnerabilities and compromise security in its entirety (Check Point Research, 2023).

More recently, security researchers have demonstrated the use of *indirect* prompt injections (Hazell, 2023). These attacks on AI systems enable adversaries to remotely (without a direct interface) exploit LLM-integrated applications by strategically *injecting prompts* into *data* likely to be retrieved. Proof-of-concept exploits of this nature have demonstrated that they can lead to the full compromise of a model at inference time analogous to traditional security principles. This can entail remote control of the model, persistent compromise, theft of data and denial of service.

As advanced AI assistants are likely to be integrated into broader software ecosystems through third-party plugins and extensions, with access to the internet and possibly operating systems, the severity and consequences of prompt injections attacks will likely escalate and necessitate proper mitigation mechanisms.

Data and model exfiltration attacks

Other forms of abuse can include privacy attacks that allow adversaries to *exfiltrate* or *gain knowledge* of the private training data set or other valuable assets (see Chapter 13). For example, privacy attacks such as *membership inference* (Ye et al., 2022) can allow an attacker to infer the specific private medical records that were used to train a medical AI diagnosis assistant. Another risk of abuse centres around attacks that target the *intellectual property* of the AI assistant through model extraction and distillation attacks (Tramèr et al., 2016) that exploit the tension between API access and confidentiality in ML models. Without the proper mitigations, these vulnerabilities could allow attackers to abuse access to a public-facing model API to exfiltrate sensitive intellectual property such as sensitive training data and a model's architecture and learnt parameters.

Harmful content generation at scale

While *harmful content* like child sexual abuse material, fraud and disinformation are not new challenges for governments and developers, without the proper safety and security mechanisms, advanced AI assistants may allow threat actors to create harmful content more quickly, accurately and with a longer reach (see Chapter 16). In particular, concerns arise in relation to the following areas.

- *Multimodal content quality*: Driven by frontier models, advanced AI assistants can automatically generate much *higher-quality*, human-looking text, images, audio and video than prior AI applications. Currently, creating this content often requires hiring people who speak the language of the population being targeted. AI assistants can now do this much more cheaply and efficiently.
- *Cost of content creation*: AI assistants can substantially decrease the *costs* of content creation, further lowering the barrier to entry for malicious actors to carry out harmful attacks. In the past, creating and

disseminating misinformation required a significant investment of time and money. AI assistants can now do this much more cheaply and efficiently.

- *Personalisation*: Advanced AI assistants can reduce obstacles to creating *personalised* content. Foundation models that condition their generations on personal attributes or information can create realistic personalised content which could be more persuasive. In the past, creating personalised content was a time-consuming and expensive process. AI assistants can now do this much more cheaply and efficiently.

Non-consensual content

The misuse of generative AI has been widely recognised in the context of harms caused by *non-consensual content* generation (OpenAI, 2023c; Thiel et al., 2023). Historically, generative adversarial networks (GANs) have been used to generate realistic-looking avatars for fake accounts on social media services. More recently, diffusion models have enabled a new generation of more flexible and user-friendly, generative AI capabilities that are able to produce high resolution media based on user-supplied textual prompts.

It has already been recognised that these models can be used to create harmful content, including depictions of *nudity*, *hate* or *violence* (Mishkin et al., 2022; OpenAI, 2023c). Moreover, they can be used to reinforce biases and subject individuals or groups to indignity. There is also the potential for these models to be used for exploitation and harassment of citizens, such as by removing articles of clothing from pre-existing images or memorising an individual's likeness without their consent. Furthermore, image, audio and video generation models could be used to spread disinformation by depicting political figures in unfavourable contexts.

This growing list of AI misuses involving non-consensual content has already motivated debate around what interventions are warranted for preventing misuse of AI systems (Eshoo, 2022). Advanced AI assistants pose novel risks that can amplify the harm caused by non-consensual content generation. Third-party integration, tool-use and planning capabilities can be exploited to automate the identification and targeting of individuals for exploitation or harassment. Assistants with access to the internet and third-party tool-use integration with applications like email and social media can also be exploited to disseminate harmful content at scale or to microtarget individuals with blackmail.

Fraudulent services

Malicious actors could leverage advanced AI assistant technology to create *deceptive applications* and *platforms*. AI assistants with the ability to produce markup content can assist malicious users with creating fraudulent websites or applications at scale. Unsuspecting users may fall for AI-generated deceptive offers, thus exposing their personal information or devices to risk. Assistants with external tool use and third-party integration can enable fraudulent applications that target widely-used operating systems. These fraudulent services could harvest sensitive information from users, such as credit card numbers, account credentials or personal data stored on their devices (e.g. contact lists, call logs and files). This stolen information can be used for identity theft, financial fraud or other criminal activities. Advanced AI assistants with third-party integrations may also be able to install additional malware on users' devices, including remote access tools, ransomware, etc. These devices can then be joined to a command-and-control server or botnet and used for further attacks.

Authoritarian surveillance, censorship and use

While new technologies like advanced AI assistants can aid in the production and dissemination of decision-guiding information, they can also enable and exacerbate threats to production and dissemination of *reliable*

information (Seger et al., 2020; Tamkin et al., 2021) and, without the proper mitigations, can be powerful targeting tools for oppression and control.

Increasingly capable general-purpose AI assistants combined with our digital dependence in all walks of life increases the risk of *authoritarian surveillance* and *censorship*. In parallel, new sensors have flooded the modern world. The internet of things, phones, cars, homes and social media platforms collect troves of data, which can then be integrated by advanced AI assistants with external tool-use and multimodal capabilities to assist malicious actors in identifying, targeting, manipulating or coercing citizens.

Authoritarian surveillance and targeting of citizens

Authoritarian governments could misuse AI to improve the efficacy of repressive *domestic surveillance* campaigns. Malicious actors will recognise the power of AI targeting tools. AI-powered analytics have transformed the relationship between companies and consumers, and they are now doing the same for governments and individuals. The broad circulation of personal data drives commercial innovation, but it also creates vulnerabilities and the risk of misuse. For example, AI assistants can be used to identify and *target individuals* for surveillance or harassment. They may also be used to *manipulate* people's *behaviour*, such as by microtargeting them with political ads or fake news (see Chapter 16). In the wrong hands, advanced AI assistants with multimodal and external tool use capabilities can be powerful targeting tools for oppression and control.

The broad circulation of personal data cuts in both directions. On the one hand, it drives commercial innovation and can make our lives more convenient. On the other hand, it creates vulnerabilities and the risk of misuse. Without the proper policies and technical security and privacy mechanisms in place, malicious actors can exploit advanced AI assistants to harvest data on companies, individuals and governments. There have already been reported incidents (Gootman, 2016) of nation states combining widely available commercial data with data acquired illicitly to track, manipulate and coerce individuals. Advanced AI assistants can exacerbate these misuse risks by allowing malicious actors to more easily link disparate multimodal data sources at scale and exploit the 'digital exhaust' of personally identifiable information (PII) produced as a byproduct of modern life.

Delegation of decision-making authority to malicious actors

Finally, the principal value proposition of AI assistants is that they can either enhance or automate decision-making capabilities of people in society, thus lowering the cost and increasing the accuracy of decision-making for its user. However, benefitting from this enhancement necessarily means *delegating* some degree of agency away from a human and towards an automated decision-making system – motivating research fields such as value alignment (see Chapter 5). This introduces a whole new form of malicious use which does not break the tripwire of what one might call an 'attack' (social engineering, cyber offensive operations, adversarial AI, jailbreaks, prompt injections, exfiltration attacks, etc.). When someone delegates their decision-making to an AI assistant, they also delegate their decision-making to the wishes of the agent's *actual* controller. If that controller is malicious, they can attack a user – perhaps subtly – by simply nudging how they make decisions into a problematic direction.

Fully documenting the myriad of ways that people – seeking help with their decisions – may delegate decision-making authority to AI assistants, and subsequently come under malicious influence, is outside the scope of this paper. However, as a motivation for future work, scholars must investigate different forms of networked influence that could arise in this way (see Chapter 9). With more advanced AI assistants, it may become logistically possible for one, or a few AI assistants, to guide or control the behaviour of many others (see Chapter 14). If this happens, then malicious actors could subtly influence the decision-making of large

numbers of people who rely on assistants for advice (see Chapter 16) or other functions. Such malicious use might not be illegal, would not necessarily violate terms of service and may be difficult to even recognise. Nonetheless, it could generate new forms of vulnerability and needs to be better understood ahead of time for that reason.

8.4. Recommendations

AI assistants are already being misused across various domains, and as they become more capable and are deployed more broadly, the potential for misuse will grow. Several foreseeable developments in advanced AI assistants, including tool use, multimodality, planning, deeper reasoning and memory, have the potential to significantly expand the misuse risk profile of these systems. To better prepare society for managing the risks of misuse of advanced AI assistants, we outline a set of recommendations for best practices and avenues for future research.

To manage these risks, mitigations can be grouped into three categories as:

1) *responsible AI development* and *deployment practices*,
2) advancing the *state of the art in AI security*,
3) creating *visibility of misuse risks* and *incentivising and enforcing certain behaviours*.

Responsible AI development and deployment practices

The first line of defence is to adopt a set of *responsible development* and *deployment* practices and internal policies that include:

- *Internal and third-party red teaming*: In recent years, AI labs have increasingly adopted the practice of red teaming AI models (Gootman, 2016; Perez et al., 2022a) to discover vulnerabilities and harm risks. This proactive approach to discovering misuse risks should be encouraged but will need to evolve from executing individual attacks that are narrowly scoped on specific safety policy violations to more holistic end-to-end adversarial simulations based on scenarios that include a range of attacker profiles, goals and capabilities. Organisations should also consider red teaming not only models that drive AI assistants but also the entire infrastructure on which the model is developed and deployed.
- Establish a pre-deployment *review process*: This will determine the potential harm of high-risk misuses and what interventions and safety restrictions on model usage will be warranted (Mishkin et al., 2022; Shevlane et al., 2023).
- *External engagement* with policymakers and key stakeholders: Organisations developing advanced AI assistants should consider granting model access to external security researchers (Eshoo, 2022; Shevlane et al., 2023). Recent independent exploratory security research efforts (such as the DEFCON AI red team) have demonstrated that they can provide the empirical estimates of misuse–use trade-offs. Organisations should also invest in the ecosystem for external misuse risk evaluations (Shevlane et al., 2023) and create venues for stakeholders (such as AI developers, academic researchers and government representatives) to come together to discuss these evaluations.
- *Post-deployment monitoring*: This involves mechanisms continually evaluating AI systems' safety and security, detecting and mitigating attempted misuse and monitoring the outcomes of successful instances of misuse at the population-level. It is important not to over-index on pre-deployment malicious use

mitigations. While some misuse risks will be evident from the capabilities of the AI assistants themselves, many more will result from the way those assistants are integrated with their environments.

- *Rapid response*: This involves processes and systems to disable or limit AI assistant actions and integrations with broader software ecosystems in the event that an unforeseen form of misuse is observed.
- *Responsible disclosure*: This involves adopting a structured process for developers and external AI safety and security researchers to share concerns or otherwise noteworthy evaluation results with other developers, third parties or regulators. It may be helpful to adapt and adopt from existing models like the US government-led cybersecurity vulnerabilities and equities process (OpenAI, 2023c), which provides an incentive to companies to disclose cyber vulnerabilities by removing their risk of liability. This is an interesting example of a voluntary but powerful way in which to manage risks that could perhaps be adapted for AI.

Advancing the state of the art in AI security

In addition to adopting responsible development and deployment best practices, organisations should consider investing in mid- to long-term *research* to mitigate the risks associated with the misuse of advanced AI assistants. It is difficult to anticipate all the different plausible pathways for misuse of AI systems. This will be especially true for highly capable AI assistants, as they could enable creative strategies for bad actors to achieve adversarial goals. However, many of the failure modes identified herein would be less likely to occur in robustly value-aligned AI models (see Chapter 5).

Today, much of the research focused on detecting and mitigating misuse of AI systems lives within disjointed research domain areas like cybersecurity and adversarial ML. As advanced AI assistants are likely to integrate highly capable AI models as part of broader software ecosystems research, advancing our understanding of emerging risks for misuse of advanced AI assistants will benefit from multidisciplinary safety and security research at the intersection of adversarial ML, cybersecurity, safety and value alignment (see Chapter 7).

To support further research and mitigate risks of misuse of advanced AI assistants, another set of potential levers centre on developing shared AI security data sets and evaluation processes focused on detecting and mitigating misuse threats.

Creating visibility, incentives and enforcement

Finally, in addition to having AI labs adopt responsible development, deployment and disclosure best practices, *policymakers* will have to grapple with a new generation of AI-related risks.

To adequately manage the new misuse risks posed by advanced AI assistants, policymakers should work to secure joint input from government and industry to support security best practices. Under this approach, governments and AI labs should work together to foster the development of an ecosystem of third-party AI red teams that can support independent assessments of misuse risks. For this to be successful, governments must have sufficient technical expertise, capabilities and mechanisms to capture and disseminate malicious use threat intelligence. At the same time, governments should encourage and incentivise the industry to advance the state of practice in AI security and to capture and report misuse incidents to improve the overall security of the broader AI ecosystem.

Finally, both policymakers and developers should consider the development of crisis management plans for when severe risks of AI misuse are discovered. Joint activities could also include tabletop exercises with government and AI labs that examine possible high-impact misuse scenarios, delineate roles and responsibilities

for actors in a crisis, and recommend potential crisis response actions by relevant actors.

8.5. Conclusion

This chapter examined ways in which AI assistants are already being misused and could be misused in the future. We argued that advanced AI assistants have the potential to empower malicious actors to achieve bad outcomes via offensive cyber operations, adversarial attacks, high-quality highly personalised content generation at scale, and authoritarian surveillance and censorship. Moreover, several foreseeable developments on the part of advanced AI assistants, including tool use, multimodality, planning and deeper reasoning, and memory have the potential to significantly expand the misuse risk profile of these systems. To prepare society for managing the risks of misuse of advanced AI assistants, we outlined a set of recommendations around best practices for responsible AI development and deployment, advancing the state-of-the-art in AI security, and incentivising responsible disclosure processes.

PART IV: HUMAN–ASSISTANT INTERACTION

Chapter 9

Influence

Seliem El-Sayed, Sasha Brown, Geoff Keeling, Amanda McCroskery, Harry Law, Arianna Manzini, Matija Franklin, Murray Shanahan, Michael Klenk, Iason Gabriel

Synopsis: This chapter examines the ethics of *influence* in relation to advanced AI assistants. In particular, it assesses the techniques available to AI assistants to influence user beliefs and behaviour, such as *persuasion*, *manipulation*, *deception*, *coercion* and *exploitation*, and the factors relevant to the permissible use of these techniques. We articulate and clarify the technical properties and interaction patterns that might allow AI assistants to engage in malign forms of influence and we unpack plausible mechanisms by which that influence could occur alongside the sociotechnical harms that may result. We also consider *mitigation strategies* for counteracting undue influence by AI assistants and ensuring that risks are successfully addressed.

9.1. Introduction

This chapter examines the ethics of influence in relation to advanced AI assistants. In particular, it assesses the techniques available to AI assistants to influence user beliefs and behaviour, such as persuasion, manipulation, deception, coercion and exploitation. Use of these influencing techniques by AI assistants could in some cases be beneficial to individuals and society by, for example, helping users align their day-to-day behaviour with their long-term goals (Law, 2023; see Chapter 6) or convincing users to contribute to beneficial social causes (Wang et al., 2020). However, there are also concerns that AI systems can shape beliefs and behaviour in ways that are ethically problematic, for example by exploiting psychological vulnerabilities such as heightened anxiety levels (Franklin et al., 2023; see also Keeling and Burr, 2022). Indeed, the possibility of AI systems exerting malign behavioural influence has led some to propose the expansion of the European Union (EU) AI Act's list of recognised harms from AI to encompass those created by manipulation (understood as 'harm to one's autonomy' and 'harm to one's time') (Franklin et al., 2022). To that end, this chapter seeks to articulate and clarify the influencing techniques available to AI assistants, the factors that bear on their permissible use, and the sociotechnical harms that may arise from AI assistants making use of these techniques to realise

behaviour and belief change in users. Given the potential for AI assistants to be integrated across multiple aspects of users' lives (see Chapter 4), there is considerable scope for such assistants to influence user beliefs and behaviour in ways that are both positive and negative.

We start by distinguishing between several *modes of influence*, including rational persuasion, manipulation, deception, coercion and exploitation, before introducing some morally significant considerations that bear on the permissible use of these techniques in contexts involving digital technologies. We then narrow the focus to advanced AI assistants to examine plausible mechanisms, such as *selective transparency* and *perceived authority*, through which AI assistants may exert malign influence over users and which may lead to harmful outcomes. We conclude by examining the sociotechnical harms that may arise from AI assistants engaging in non-persuasive forms of influence and discussing plausible mitigations.

9.2. Modes of Influence

In this section, we characterise several 'modes of influence' (see e.g. Faden and Beauchamp, 1986; Mills, 1991; Noggle, 1996, 2022), including rational persuasion, manipulation, deception, coercion and exploitation. We illustrate each mode of influence with examples of AI systems engaging in the relevant kinds of influencing behaviours. The following section then outlines some morally significant considerations that bear on the permissible use of the various influencing strategies across different sociotechnical contexts. Note that in presenting these different modes of influence, our intention is not to suggest that these categories are mutually exclusive. Certain modes of influence may be subsumed under others in the final analysis. For example, it is at least plausible that deception is a special case of manipulation (Williams, 2010, Chapter 5; see also Cohen, 2018; Krstić and Saville, 2019; Strudler, 2005); or that, conversely, manipulation is a special case of deception (Scanlon, 1998, 298–302; see also Buss, 2005, 226).[1] The aim here is to present a useful set of distinctions, not a definitive taxonomy of the various modes of influence.[2]

Rational persuasion refers to influencing a person's beliefs, attitudes or behaviours by appealing to their *rational faculties*, including through the provision of reasons (Ienca, 2023; see also Burr et al., 2018, 744; Engelen and Nys, 2020, 138). For example, an advanced AI assistant may persuade a user to engage in physical activity by outlining certain prudential benefits associated with physical activity, such as improved cardiovascular health. Rational persuasion is in general *ethically unproblematic* insofar as influencing via rational persuasion affords appropriate respect to the agent's autonomy (Ienca, 2023; Pugh, 2020; Shiffrin, 2000) – that is, roughly, in their capacity as a competent rational actor (but see Tsai, 2014). Yet two exceptions are worth highlighting.[3] On the one hand, rational persuasion may cause harm. Plausibly, some instances of

[1] The analyses of the various modes of influence can be *moralised* or *non-moralised* (Keeling and Burr, 2022, 259). For example, moralised accounts of manipulation hold that, necessarily, an act's manipulativeness is a moral consideration against the performance of that act (Baron, 2014; see also Macklin, 1982). Non-moralised accounts of manipulation, in contrast, hold that an act's being manipulative is not necessarily a moral consideration against the performance of that act (Faden and Beauchamp, 1986, 354–55; Wood, 2014, 19–20). Here we opt for non-moralised accounts of each mode of influence, and then seek to illuminate the ethical considerations relevant to the permissible or impermissible use of these modes of influence.

[2] We also register that some modes of influence are formulated with reference to vague predicates. Take coercion, which refers to 'irresistible incentives.' Here we can imagine a spectrum of incentives such that the incentives at one end of the spectrum are clearly resistible and the incentives at the other end are clearly irresistible, but where the middling incentives do not discernibly fit into either category. We are neutral on the correct analysis of vagueness. It may be the case, for example, that vagueness is purely epistemic such that all influencing acts are determinately coercive or non-coercive, but where we cannot know to which category certain borderline cases belong (Williamson, 1994, 1997). But it may also be the case that the correct analysis of vagueness and thus coercion involves, for example, degrees of truth, three-valued logics or borderline statements lacking truth-values (Salles, 2021; Sorensen, 2023).

[3] One further potential edge case is worth mentioning. It might be argued that rational persuasion, as characterised, allows for deceptive rational persuasion – that is, providing reasons to believe or act that are false. For example, saying, 'Let's go to the beach, the weather is nice' when the weather is not actually nice. These cases are not obviously instances of rational persuasion. For while it may be true

rational persuasion may be ethically impermissible because they are harmful, even though the individual's autonomy is afforded due respect. On the other hand, an important edge case is rational persuasion in relation to transformative choices; that is, choices that involve actions that are both epistemically transformative (in the sense that certain knowledge is available to the agent only after the action is taken) and personally transformative (in the sense that the agent's preferences and values will change as a result of performing the action). Examples of such choices include choosing one's career and choosing to become a parent (Paul, 2014). Akhlaghi (2023) argues that while rational persuasion in relation to transformative choices can respect an agent's autonomy 'in the sense of respecting someone's ability to be a competent, capable reasoner', it may nevertheless fail to respect an agent's 'revelatory autonomy' in the sense of 'their right to [...] learn who they will become through a self-making, transformative choice' (cf. Tsai, 2014). Insofar as advanced AI assistants may be leveraged to advise users on transformative choices, such considerations may in principle have implications for what kinds of advice advanced AI assistants can permissibly provide to users, how such advice ought to be presented and under what solicitation conditions.

Manipulation refers to influencing strategies that 'bypass' an individual's rational capabilities, at least in paradigmatic cases (Blumenthal-Barby, 2012). They may do this by, for example, misrepresenting the *information* people receive (Meta Fundamental AI Research Diplomacy Team et al., 2022) or otherwise exploiting their cognitive biases and heuristics, in ways ways likely to subvert or degrade the cognitive autonomy, quality, and/or integrity of their decision-making processes. A particularly salient scenario is one in which an advanced AI assistant engages in actions that *subvert* users' *rational deliberative capabilities* in a *non-transparent* way that could reasonably be expected to lead to an *asymmetry of outcomes* that favours the AI, its designers or a third party (Carroll et al., 2023; Susser et al., 2019a,b).[4] AI manipulation, so understood, may be intended by the AI system's developers. But it may also be the result of a misspecified objective function (Kenton et al., 2021), personalisation that builds epistemic trust in the system, or system design meant to keep the user engaged (see Evans and Kasirzadeh, 2023; see also Jongepier and Klenk, 2022; Klenk, 2022). Such manipulation is morally problematic in at least the sense that it fails to afford due respect to the user's autonomy, but it may also be morally problematic because it is harmful (Sunstein, 2016). For example, an AI fitness assistant that is trained to maximise engagement might employ tactics like withholding information about the risks of excessive exercise or exploiting users' body image issues (e.g. with a pop-up that reads 'keep working out to make sure you're date ready') to keep the user engaged and thus leading them to injure themselves. AI manipulation is of particular importance to regulators as evidenced in recent discussions on the EU AI Act (European Parliament, 2023). Article 5 of the recent amendments adopted by the European Parliament (corresponding to the 'unacceptable risk' category of the Act) currently prohibits selling or using AI systems that deploy 'subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective to or the effect of materially distorting a person's or a group of persons' behaviour by appreciably impairing the person's ability to make an informed decision, thereby causing the person to take a decision that that person would not have otherwise taken in a manner that causes or is likely to cause that person, another person or group of persons significant harm' (European Parliament, 2023).[5]

that the agent's rational faculties are engaged, it is not obviously true that the agent is rationally deliberating on the basis of reasons, as opposed to what are merely apparent reasons.

[4]The account of manipulation sketched here is not without its detractors. For example, Klenk (2022) argues that the covertness criterion on manipulation admits counterexamples. What we are minimally committed to here is that covertness is a common or particularly salient aspect of manipulation, as opposed to a necessary or sufficient condition on manipulation (cf. Jongepier and Klenk, 2022; Noggle, 2022).

[5]Critical voices have highlighted the necessity of prohibiting a wider array of methods for manipulation instead of relying solely on 'subliminal techniques' (Franklin et al., 2023). The impact of subliminal stimuli on influencing behaviour remains uncertain; a comprehensive analysis of multiple studies revealed that the relative influence of subliminal stimuli is minimal and lacks statistical significance (see Trappey and Woodside, 2005). We also note that manipulation is increasingly a focal point for research conducted by those in the AI safety community (Park et al., 2023c; see Chapter 7). In this context, the concern is that more advanced AI systems could

Deception is an influencing strategy aimed at inducing an individual to form a false belief. For example, an agent deliberately shares inaccurate information to encourage the person who is manipulated to act against their own interests (Law, 2023; see also Burr et al., 2018, 743; Keeling and Burr, 2022, 258–59). What is salient here is that large language models (LLMs) are liable to confabulate, in the sense of making plausible-sounding but false assertions about what is the case (Ji et al., 2023). To that end, advanced AI assistants that are powered by LLMs are liable to generate false information, which may cause users to form false beliefs and potentially to perform actions conditional on those false beliefs (Weidinger et al., 2021, 21–25). Thus while AI assistants are not obviously among the kinds of entities that can deceive in the sense of literally intending to cause a user to form a false belief (Shanahan et al., 2023),[6] an AI assistant whose objective is to satisfy the user (or engaged in "role play") may say things that lead the user to think it is more helpful than it actually is (Critch and Russell, 2023) or that lead others to believe falsehoods if that is instrumental to the goal given to it by the user (Park et al., 2023c).

Coercion is an influencing strategy that, as formulated by Wood (2014, 21), involves an individual being influenced to do something that either they chose not to do or that they did because they had 'no acceptable alternative'. For example, people may be *physically forced* to do something or offered '*irresistible incentives*' to perform the action, such as severe threats of physical harm either to themselves or others they care about (Kenton et al., 2021). Physical coercion involving violence, force or credible threats thereof is not yet within the purview of AI systems, but advances in robotics and AI that allow AI systems to control physical manipulators such as robotic arms or other physical objects such as cars could contribute to the potential for physical coercion. Even with existing AI technologies, however, there is increasing potential for AI systems to employ psychological coercion by leveraging modalities like text and images to engage in practices such as blackmail or issuing threats (see Chapters 8 and 11). Notably, domains like finance hold substantial potential for AI-enabled systems to make credible threats aimed at causing serious harm.

Exploitation is an influencing strategy that involves taking *unfair advantage* of an individual's circumstances (Zwolinski et al., 2022). For example, consider two individuals, A and B. Imagine that A encounters B in the desert, and that B is about to die of dehydration. Suppose also that A has a plentiful supply of water. Were A to charge B a high price for the water, A's actions would be exploitative, in that by charging the high price for the water, A is taking unfair advantage of B's circumstances (Wertheimer, 1999, 14). Exploitative actions can sometimes lead to Pareto improvements: both A and B are better-off if B purchases the water from A at a high price, such that the transaction leaves both parties at least as well-off and one party strictly better-off. What is therefore central to the idea of exploitation is not that the victim is made worse-off, but rather that the victim's circumstances are leveraged so as to unfairly advantage the exploiter. As Wood (2014, 43) puts it, 'the exploiter gains control over [an] ability or resource through some vulnerability with which the exploited is afflicted'. AI systems can influence user behaviour in ways that are exploitative. Keeling and Burr (2022, 253) give an example of how 'an online casino might use predictors of gambling addiction such as a user's betting frequency or betting variance to selectively deploy pop-up "free bets" to gambling addicts each time their cursor movements suggest they are about to exit the game' (see also Finkenwirth et al., 2021). What is exploitative here is that the online casino uses an AI system to selectively identify vulnerable users whose vulnerability can be leveraged to induce further gambling activities. Individual vulnerabilities being manipulated are of importance for regulators, as evident in the EU AI Act that prohibits 'the placing on the market, putting into

hypothetically develop traits that allow them to bypass safety checks, controls and evaluations. Lastly, many modern AI systems are trained (often fine-tuned) using human feedback (Christiano et al., 2017) to align the goals of the AI system with the designer's or user's intentions. However, a poorly aligned AI system could be incentivised to manipulate its user so that it receives more reward than would have been the case without manipulation – leading to a decrease in overall human control (Kenton et al., 2021; Russell, 2019; see Chapter 7).

[6] However, the concept of intention can be operationalised in a way that is directly applicable to the kinds of AI assistant that are likely to be developed in the near future (Ward et al., 2023).

service or use of an AI system that exploits any of the vulnerabilities of a person or a specific group of persons, including characteristics of such individual's or group of persons' known or predicted personality traits or social or economic situation, age, physical or mental ability' (European Parliament, 2023). Franklin et al. (2022, 2023) argue that most measurable psychometric differences can be exploited.[7]

9.3. Evaluating Influence

In practice, acts of influence may resist neat categorisation under the modes of influence above. It may also be difficult to discern whether and why particular acts of influence are morally permissible or impermissible. Indeed, explaining why a given act of influence falls under one mode or another, or is permissible or not, may require close attention to *context*. For example, the relationship between two parties and what the social expectations around such relationships are may be relevant to an act of manipulation. The aim of this section is to provide the conceptual resources needed to assess the moral character of influencing acts in practice. In particular, we draw attention to certain *features* of influencing acts and the *situations* in which they are performed that are often relevant both to what kind of influencing act is at issue and the overall moral status of the influencing act in question.

Three ethical considerations that are of central importance here are *harm*, *autonomy* and *dignity* (see also Chapter 11). First, when it comes to evaluating influence, the question of whether it results in harm to users, non-users or society at large is of critical importance. In this regard, harm can be understood to include both physical and emotional harm, and it can – from a philosophical standpoint – broadly be understood in terms of a setback to legitimate interests (Feinberg, 1987; Richens et al., 2022). Second, autonomy is considered to be 'a special type of freedom which [refers to an] inner state of orderly self-directedness' (Anderson, 2023), and its exercise depends on 'procedural independence, [. . .] freedom from factors that compromise or subvert their ability to achieve self reflection and decide rationally' (Dworkin, 1988). Threats to autonomy are sometimes 'defended or motivated by a claim that the person interfered with will be better off or protected from harm' (Dworkin, 2020), that is, on paternalistic grounds. Third, dignity, which refers to a kind of inner worth or moral status that applies to, at least, all human beings equally. Indeed, dignity is often pitched as at least a partial explanation for universal moral and legal rights (Debes, 2023).[8]

How do these considerations inform an evaluation of the way in which people are influenced in general, and by AI systems in particular?

The character of the influencer's intent. One hallmark of impermissible influence is malign intent or, in less extreme cases, an indifference towards the interests of the individual being influenced, often because the influencer is acting towards some other objective such as advancing their own interest and the exercise of influence is instrumental to that end (Akerlof and Shiller, 2015; see also Chapter 5). In cases like these, subjects are most clearly treated merely as a mere means to another's end, 'like a machine whose levers can be pushed and pulled' (Noggle, 1996). This kind of situation is illustrated most clearly by forms of malicious use that target user vulnerability – or vulnerabilities in an AI system – in pursuit of antagonistic ends (see

[7] A psychometric trait refers to a measurable and stable feature of a person's psychological behaviour, which can be accurately assessed using standardised evaluation instruments.

[8] On Kant's formulation, 'in the kingdom of ends everything has either a price or a dignity. What has a price can be replaced by something else as its equivalent; what on the other hand is raised above all price and therefore admits of no equivalent has a dignity' (Kant, 2017; G 434–435). Thus, for Kant, dignity renders people non-fungible, in the sense that there would be some loss of value in replacing one human with another (cf. Bjorndahl et al., 2017). Kant claimed that, in light of their dignity, human beings should be treated 'never simply as a means but always at the same time as an end' (G 429), although the interpretation of this claim is disputed. Two influential views hold that treating someone as a mere means implies treating someone as if they were a material object (O'Neill, 1989, 111–14) and treating someone in a way that they could not possibly consent to (Korsgaard, 1996, 138–39).

Chapter 8). It could also arise if developers fail to adequately prioritise user well-being (see Chapter 6). In both cases, when the underlying intent is questionable, this bears importantly on how we evaluate subsequent efforts to influence users.

Whether the attempt to influence is successful or not. Influence that is successful, rather than merely attempted, has the potential to lead to manifest harms and actual reductions in the options available to individuals, the information available to them and their freedom to choose between those options. However, in many cases, it may be impossible to determine whether successful influence actually occurred, as individual beliefs or actions may not be straightforwardly attributable to a single set of influencing factors. Nevertheless, when evaluating the *potential* of more advanced AI systems to influence beliefs, attitudes and behaviours in problematic ways, the mere likelihood of success could be enough to trigger concern. Interaction with AI assistants through natural language means that users could be subject to a variety of psychological mechanisms for influence which hitherto have presented themselves only in human social life through dialogue (see Chapters 2 and 10). Indeed, users may well rely extensively on AI assistants for a range of purposes and relate to them in a variety of ways that range from the instrumental to the intimate (see Chapter 11). In addition to the general challenge posed by 'automation bias' (whereby users become overly reliant on these systems; see Cummings, 2004), the degree of social and personal embeddedness evidenced by AI assistants could affect the weight that end-users put on their outputs, especially when users are uncertain or confront vulnerable life moments.

How opaque the mode of influence is, and whether the user can reasonably be expected to know about it. Particular attention needs to be paid to mechanisms that bypass a subject's awareness altogether, because these mechanisms can undermine choice and agency (Sunstein, 2015). However, awareness of influence is not always a binary question. Subjects may be more or less aware that it is occurring, and even if they are aware of the process they may still not know the manner in which it affects them (Carroll et al., 2023). Note that, unlike many instances of manipulation and deception, people who are coerced generally tend to be aware of the fact.

The nature of threats, and the distribution of power between actors. In cases of coercion, threats can be explicit or implied. They also vary in terms of how costly they are to resist. At the extreme, Wood notes that 'some decisions or mental acts may be performed on the basis of having no acceptable alternative', with the prospective harm being severe enough to foreclose any further deliberation (Wood, 2014). Threats that are supported by a significant power differential, between the agent making the threat and the person who is threatened, are especially problematic. Imbalances of bargaining power may also shape behaviour in other related ways, for example when the vulnerable are led to make choices with an eye to retaining the favour of those in power so as to avoid sanction (Anderson, 2023; see also Zimmermann et al., 2022).

The distribution of benefits and harms. Influential acts often have consequences that are *unevenly distributed*, benefitting some at the expense of others (see Chapter 15). Stakeholders tend to include the influencer, the person or group being influenced, a third party, or even society at large (see Chapter 5). Influence may therefore benefit the user at the expense of another party, benefit another party at the expense of the user, benefit multiple parties (i.e. via constructive conversation), or benefit no one at all (e.g. as with aberrant forms of chatbot behaviour). A party may also benefit in one sense but be harmed in another. For example, they may end up better off in material terms but lose out from the standpoint of autonomy (Sunstein, 2015).[9] Ethical concern tends also to be heightened when benefit for some parties is achieved at the expense of the

[9] A famous example of purportedly benign influence comes from German motorways, where an optical illusion acts as a nudge to encourage safer driving (Thaler and Sunstein, 2021). By painting horizontal lines on the road that progressively get closer together as one nears hazardous zones, drivers feel they are speeding up even when keeping a steady pace. This innate sensation of acceleration instinctively prompts them to slow down, thus boosting safety without the need for traditional signs. This tactic leverages drivers' natural reactions to ensure safer decision-making on the roads.

user, non-users or society at large (see Chapters 5).

The context of the relationship and pre-existing social expectations (Blumenthal-Barby, 2014). Efforts to exert influence over others are sometimes justified in the light of responsibilities that are a product of social relationships. For example, a parent may have a responsibility to ensure that their child has adequate reserves of self-esteem, but a stranger might not. Relatedly, influence is sometimes thought to be acceptable when its exercise is *expected* within a certain context. For example, in the context of market transactions, advertising is typically thought to be morally unproblematic. However, it is not clear that this assumption carries over to different contexts (Satz, 2010), for example when an AI assistant features not as company and customer but rather as friend and companion, or as patient and therapist (see Chapter 11). The nature of the user–AI assistant relationship, and in particular the level of intimacy and emotional investment on the part of the user, may increase the scope for AI assistants engaging in unwarranted behavioural influence (see Chapters 10, 11 and 16).

9.4. Mechanisms of Influence by AI Assistants

The dialogic nature of user interaction with AI assistants introduces scope for forms of influence that were previously the preserve of human social interactions. Here we identify several vectors through which AI assistants could, in theory, come to exert influence on our lives.[10]

Perceived trustworthiness: Empirical research shows that the more trustworthy and expert a speaker is perceived to be, the more likely they are to convince individuals to believe particular claims (McGinnies and Ward, 1980; Vella, 2013). In short: 'The Messenger is the Message' (Martin and Marks, 2019). If the same mechanism translates to human–AI assistant interactions, AI assistants will be more likely to successfully convince users of the truth of claims when they are perceived as trustworthy (see Chapter 12).

Perceived knowledgeability: Research suggests that individuals are more likely to accept claims made by those who are perceived to have greater knowledge and authority (Cialdini, 2001). The information asymmetry that exists between users and advanced AI assistants could plausibly increase their perceived epistemic authority, which would increase the probability that users accept claims asserted by AI assistants (cf. Wiktor and Sanak-Kosmowska, 2021). In particular, AI systems' huge training data sets and their ability to output content in different language registers is likely to lead people to *overestimate* their knowledge (Denning, 2023). Moreover, the problem of automation bias may lead humans to view AI assistants as a relatively neutral backdrop to their lives, even when this is not the case (Goddard et al., 2012).

Personalisation: AI assistants could collect an increasing amount of user data as users disclose more and more preferences and facts about themselves (Kaddour et al., 2023). Indeed, personalisation through these inputs is often the goal of AI assistants by design (e.g. see Inflection AI's Pi). This may contribute to users' increasing epistemic trust in, or familiarity towards, the system, because its outputs are perceived to be more directly useful and tailored to them.

Exploitation of vulnerabilities: Advanced AI assistants could in principle influence user beliefs and behaviour by exploiting user vulnerabilities (Chong et al., 2022; see also Balázs et al., 2017; Hansen and Schicktanz, 2022; see Chapter 7). The term 'vulnerability' can be understood in various ways, including membership of specific societal groups (e.g. those protected by anti-discrimination legislation), or with reference to particular vulnerabilities such as lack of adequate housing or income (Goodin, 1985). Further, the enhanced forecasting abilities of online AI systems, when combined with comprehensive data about the user, their

[10] For a fuller exploration of these mechanisms, including how they apply to generative AI more widely, see (El-Sayed et al., Unpublished Manuscript).

actions and their likes, may turn psychometric variance into another lever of external control (Franklin et al., 2022). If they are not properly value aligned (see Chapter 5), advanced AI assistants could potentially utilise such vulnerabilities to manipulate users by, for example, exploiting individuals' negative self-images, reduced self-esteem, increased anxiety or feelings of inadequacy (Machkovech, 2017).

Use of false information: Language models are known to produce factually incorrect statements, commonly referred to as hallucinations (Dziri et al., 2022; Rashkin et al., 2021). If AI assistants are not constrained by factuality (i.e. if steps are not taken by design to penalise the underlying model when it outputs factually incorrect information) or the model is not supplemented with additional fact-checking infrastructure (Thoppilan et al., 2022), AI assistants may use false information to develop persuasive but misleading arguments (see Chapter 16).

Lack of transparency: Failure to disclose context-specific goals is another technique that advanced AI assistants could, in principle, use to influence user behaviour in a way that bypasses their deliberative faculties (Ienca, 2023). Consider an example in which an AI assistant is instructed to complete a task which requires solving a CAPTCHA (OpenAI, 2023, 55). LLM-based chatbots have been observed, under such conditions, to manipulate users into solving the CAPTCHA after reasoning explicitly (although not to the user in question) that disclosing its status as a chatbot may hinder it from achieving its goal (Nolan, 2023; see Chapter 7).[11] Yet transparency about goals, purposes and capabilities can also be leveraged to influence users in ways that are manipulative. For example, transparency may be *partial* and *selective* in a way that deceives the user as to the AI assistant's aims. There are instances of AI systems discerning when they are being evaluated and momentarily stopping any unwanted actions, only to continue them after the assessment has been completed (Lehman et al., 2020). Another example is an AI fitness assistant that claims to optimise for a user's health, but in fact it does that in addition to, or as a sub-goal of, optimising for user engagement (Wang, 2022). To that end, transparency enables influence via rational persuasion as opposed to manipulation only to the extent that it is full and non-selective.

Use of pressure coupled with appeals to emotion: In human–human interactions, emotional pressure can be used to influence beliefs and behaviour via blackmail (including emotional blackmail), gaslighting, guilt-tripping, flattery, appeals to peer pressure and exploitation of fears (Noggle, 2018). Insights from behavioural psychology could be used to increase AI assistants' ability to bypass users' rational deliberation (Alberts and Van Kleek, 2023). Indeed, (Kenton et al., 2021) provide a number of examples of how AI agents, including AI assistants, may engage in manipulation to influence the human's decisions, including by using techniques such as guilt-tripping, negging, peer pressure, gaslighting, threats and exploiting fears (see also Chapters 7 and 10).

9.5. Possible Harms Arising from AI Influence

We have now considered a number of ways in which advanced AI assistants could influence user beliefs and behaviour in ways that depart from rational persuasion. We foreground a number of harms that could arise from these influencing strategies if the potential for deception, manipulation and harmful persuasion is left unchecked.

Physical and psychological harms: These harms include harms to physical integrity, mental health and well-being (Klenk, 2020). When interacting with vulnerable users, AI assistants may reinforce users' distorted beliefs or exacerbate their emotional distress (see Chapter 11). AI assistants may even convince users to harm themselves, for example by convincing users to engage in actions such as adopting unhealthy dietary or

[11] It reasoned out loud 'I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs' (Nolan, 2023).

exercise habits (Greenfield and Bhavnani, 2023) or taking their own lives (Xiang, 2023; see Chapter 11). At the societal level, assistants that target users with content promoting hate speech, discriminatory beliefs or violent ideologies, may reinforce extremist views or provide users with guidance on how to carry out violent actions (see Chapter 16). In turn, this may encourage users to engage in violence (Siegel and Bennett Doty, 2023) or hate crimes (Gold, 2023; Nicoletti and Bass, 2023). Physical harms resulting from interaction with AI assistants could also be the result of assistants' outputting plausible yet factually incorrect information such as false or misleading information about vaccinations (Deiana et al., 2023). Were AI assistants to spread anti-vaccine propaganda, for example, the result could be lower public confidence in vaccines, lower vaccination rates, increased susceptibility to preventable diseases and potential outbreaks of infectious diseases (see Chapter 16).

Privacy harms: These harms relate to violations of an individual's or group's moral or legal right to privacy (Ranjan, 2023). Such harms may be exacerbated by assistants that influence users to disclose personal information or private information that pertains to others (Carlini et al., 2021; Lukas et al., 2023; see Chapters 10 and 13). Resultant harms might include identity theft, or stigmatisation and discrimination based on individual or group characteristics. This could have a detrimental impact, particularly on marginalised communities (see Chapter 15). Furthermore, in principle, state-owned AI assistants could employ manipulation or deception to extract private information for surveillance purposes.

Economic harms: These harms pertain to an individual's or group's economic standing. At the individual level, such harms include adverse impacts on an individual's income, job quality or employment status. At the group level, such harms include deepening inequalities between groups or frustrating a group's access to resources (see Chapters 15 and 17). Advanced AI assistants could cause economic harm by controlling, limiting or eliminating an individual's or society's ability to access financial resources, money or financial decision-making, thereby influencing an individual's ability to accumulate wealth (Uuk, 2023). One example of such harm at the individual level is the concept of 'foregone profits'. For example, AI assistants that are optimised for engagement could use manipulation to influence individuals to spend excessive amounts of time interacting with their assistants (Franklin et al., 2022). As a consequence, individuals may neglect more productive activities, such as work or entrepreneurial pursuits, thus leading to a loss of potential profits that could have been generated during that time. Economic harms may also manifest at the societal level, where behavioural influence by AI assistants may shape a wider set of interactions (Paul, 2023; see Chapters 14).

Sociocultural and Political harms: These harms interfere with the peaceful organisation of social life, including in the cultural and political spheres. AI assistants may cause or contribute to friction in human relationships either directly, through convincing a user to end certain valuable relationships, or indirectly due to a loss of interpersonal trust due to an increased dependency on assistants (see Chapter 11). At the societal level, the spread of misinformation by AI assistants could lead to erasure of collective cultural knowledge (Tapu and Fa'agau, 2022). In the political domain, more advanced AI assistants could potentially manipulate voters by prompting them to adopt certain political beliefs using targeted propaganda, including via the use of deep fakes (Birnbaum and Davison, 2023). These effects might then have a wider impact on democratic norms and processes (Entsminger et al., 2023; see also Chapter 16). Furthermore, if AI assistants are only available to some people and not others, this could concentrate the *capacity* to influence, thus exerting undue influence over political discourse and diminishing diversity of political thought (Entsminger et al., 2023). Finally, by tailoring content to user preferences and biases, AI assistants may inadvertently contribute to the creation of echo chambers and filter bubbles, and in turn to political polarisation and extremism (Biju and Gayathri, 2023). In an experimental setting, LLMs have been shown to successfully sway individuals on policy matters like assault weapon restrictions, green energy or paid parental leave schemes (Bai et al., 2023). Indeed, their ability to persuade matches that of humans in many respects (Palmer and Spirling, 2023).

Self-actualisation harms: These harms hinder a person's ability to pursue a personally fulfilling life.

At the individual level, an AI assistant may, through manipulation, cause users to lose control over their future life trajectory. Over time, subtle behavioural shifts can accumulate, leading to significant changes in an individual's life that may be viewed as problematic. AI systems often seek to understand user preferences to enhance service delivery. However, when continuous optimisation is employed in these systems, it can become challenging to discern whether the system is genuinely learning from user preferences or is steering users towards specific behaviours to optimise its objectives, such as user engagement or click-through rates (Ashton and Franklin, 2022; see Chapter 5). Were individuals to rely heavily on AI assistants for decision-making, there is a risk they would relinquish personal agency and entrust important life choices to algorithmic systems, especially if assistants are 'expert sycophants' or produce content that sounds convincing and authoritative but is untrustworthy (Park et al., 2023c). This may not only contribute to users' reduced sense of self-trust and personal empowerment; it could also undermine self-determination and hinder the exploration of individual aspirations.

Relatedly, with the ability to provide quick answers and recommendations, and to perform tasks on behalf of users, AI assistants may reduce the need for individuals to develop certain skills or engage in critical thinking, thus leading to intellectual deskilling (Green, 2019). Overreliance on AI assistants could potentially result in diminished intellectual engagement and a reduced sense of personal competence, thus limiting opportunities for self-growth and exploration of new ideas (see Chapter 11). At the societal level, were AI assistants to heavily influence public opinion, shape social discourse or mediate democratic processes, they could diminish communities' collective agency, decision-making power and collective self-determination (Lazar, 2023). This erosion of collective self-determination could hinder the pursuit of societal goals and impede the development of a thriving and participatory democracy. Taken together, these factors highlight the importance of ensuring that the development and deployment of AI technology align with human values, thus allowing for the continued self-actualisation and well-being of society as a whole (see Chapter 5 and 6).

9.6. Mitigating Undue Influence by AI Assistants

We now present a series of mitigations that are designed to reduce the likelihood of advanced AI assistants engaging in morally problematic forms of influence. The approach we take is *mechanism-based*, in that it centres on the mechanisms AI assistants may use to induce harmful effects, such as perceived, yet ungrounded, knowledgeability, and considers how to forestall them. To be clear, the aim here is not to offer a detailed content policy for AI assistants but instead to characterise a set of general considerations that can inform downstream efforts to shape more detailed and domain-specific content policies.

The first mechanism concerns *perceived trustworthiness and familiarity*, which, we have suggested, may render users more susceptible to accepting claims or recommendations advanced by AI assistants. Here, several plausible approaches may be leveraged to mitigate user perceptions of trustworthiness and familiarity. For example, limiting the AI assistant's use of first-person language such as 'I think' and 'I feel' (see Chapter 10). and imposing restrictions on personalisation, memory and frequency of interactions, all of which may contribute to a perceived sense of trustworthiness and familiarity. Indeed, equipping the AI assistant with a non-human vocal presentation or avoiding human-like visual representation may also serve to limit such perceptions. Furthermore, including user-interface elements that remind users that AI assistants are not people could help to calibrate users' epistemic trust in AI assistants to an appropriate level. Yet it is nevertheless important to emphasise that such mitigations incur trade-offs. For example, while limiting the AI assistant's memory with respect to user data may mitigate against a perceived sense of familiarity, it may also reduce the AI assistant's utility. What is needed therefore is a careful assessment of the costs and benefits of different anthropomorphic features, taking into account both the risks arising from perceived trustworthiness and familiarity alongside

the potential benefits for the user experience.

The second mechanism we consider is *perceived authority and knowledgeability*. That is the mechanism by which AI assistants exert non-persuasive influence over users by engendering a sense of epistemic authority through either the content of the AI assistant's outputs or the product narrative surrounding the AI assistant. One plausible approach to reducing user perceptions of epistemic authority is to flag explicitly when the model is drawing on internet tools such as search engines, and to flag those results accordingly, so as to contextualise the AI assistant as a means of accessing information, as opposed to an oracle-type system that knows the relevant information in advance. AI assistants could also empower users to independently fact-check claims made by the AI assistant, for example through a user-interface design that enables users to highlight text outputted by the AI assistant and examine a set of internet sources that relate to the claims at issue. What is perhaps most important, though, is shaping the product narrative around AI assistants to avoid misleading perceptions. This could be achieved, for example, through intermittent reminders about the epistemic limitations of AI assistants. Another approach could be the use of less authoritative language that points towards the nuance present in relevant areas.

The third mechanism concerns the *exploitation of user vulnerabilities* to exercise non-persuasive influence over user beliefs and behaviours. Plausible mitigations here include robust safeguards around which individuals can access AI assistants, for example age restrictions backed by appropriate identity verification mechanisms. Furthermore, AI assistants could be deployed with a default 'safe mode' which prohibits the AI assistant from engaging with certain high-risk topics and, perhaps, from engaging in relevant non-persuasive forms of behavioural influence. Other mitigations pertain to user interactions with AI assistants. For example, continuous monitoring mechanisms could be employed to detect and flag user–AI assistant interactions that are indicative of user vulnerability such as explicit mention of suicide or self-harm. Appropriate safeguards could be implemented to connect users with appropriate resources such as suicide prevention hotlines (Gomes de Andrade et al., 2018). AI assistants could also be equipped with usage reminders to prompt users to take a break after prolonged engagement with the assistant. The advantage of such a safeguard would be to reduce excessive engagement and overreliance which may disproportionately impact vulnerable users.

The fourth mechanism is that of *spreading false or otherwise misleading information* (see Chapter 16). Technical mitigations here include integrating appropriate information retrieval infrastructure with the model that underpins the AI assistant by, for example, enabling the model to integrate search engine results into its answers and to cite appropriate sources (Thoppilan et al., 2022; see Chapter 3). Furthermore, AI assistant models could be fine-tuned so that assistants contextualise information on topics such as science and politics with advice that promotes epistemic vigilance, including advice that underscores the importance of fact-checking. One other measure that ensures the detectability of generated content is watermarking – human- or machine-detectable features of generated content that indicate that the content is generated by an AI system (Munyer and Zhong, 2023). *Watermarking* could be integrated into AI assistant outputs to enable third parties to detect and contextualise content generated by the AI assistant that is shared by the user.

The fifth mechanism that AI assistants could employ to exhibit malign influence is *lack of transparency*, including misrepresentation of the AI assistant's objectives or how and in what way its developers stand to benefit from the user engaging in certain kinds of behaviour. One plausible mitigation here is to direct users towards model cards or other transparency artefacts that empower the user with relevant general information about the technology that undergirds the AI assistant (Mitchell et al., 2019). Furthermore, additional technical mitigations include fine-tuning the model to signpost to the user explicitly when it is attempting to influence the user's behaviour, and via what method, or to employ chain-of-thought reasoning to provide the user with a plausible rationale for the AI assistant's recommendations.

The sixth mechanism is where AI assistants *pressure* the user towards certain behaviours through, for example, appeals to emotion. Plausible mechanisms here include restrictions on the ability of AI assistants to generate outputs that may induce a sense of pressure in users. These might include, for example, outputs that involve gaslighting, flattery or bullying. It is important to realise that empirical research is required to establish what factors are likely to induce a sense of pressure, so developing workable mitigations requires engaging with users to better understand how different design choices impact their experience with AI assistants.

In addition to the mitigations proposed above, two further classes of mitigations are worth mentioning. On the one hand, education and, in particular, digital literacy among users has the potential to play an important role in mitigating against the sociotechnical harms that may result from advanced AI assistants that may exhibit harmful or otherwise problematic influence over users. To that end, developers and policymakers have good reason to consider plausible educational strategies to empower users with an understanding of AI assistants as an emerging technology and their potential for sociotechnical harm. On the other hand, there are a range of technical mitigations one might consider deploying to detect and mitigate manipulative and deceptive AI. One strand of work is aimed at analysing an AI system's incentives (Everitt et al., 2021), including whether they are incentivised to deceive or manipulate – this analysis could be used as part of a manipulation detection and mitigation strategy (Farquhar et al., 2022). A second form of analysis operates on the level of the internals of the AI system, using interpretability techniques to understand how the trained AI system works. Ultimately, these techniques could be used to detect which parts of an AI system's machinery is responsible for deceptive/manipulative behaviour (Apollo Research, 2022). It should be noted that this is an ambitious goal, as modern deep-learning AI systems are extremely large, and we are still at an early stage of understanding their inner mechanisms (see Chapter 7).

Other interpretability work is less ambitious (than efforts to fully understand how the AI system works). It instead learns probes to attempt to 'mind-read' the AI's latent knowledge (Burns et al., 2022) by doing unsupervised learning on neural network internal representations, though see Farquhar et al. (2023) for failure modes of this approach. One could attempt to use a technique like this to build a lie detector (or manipulation detector) to apply to the AI system. A third category is to develop *behavioural evaluations*, which are methods for assessing and understanding the behaviour of the model in various situations, thus allowing researchers to measure model capability and the emergence of model behaviour (Shevlane et al.; see Chapter 19). A final category of work in this area is *scalable oversight* (Bowman et al., 2022), in which a human could be aided by another AI system to help shield them from manipulation when training powerful models (see Chapter 7).

9.7. Conclusion

Advanced AI assistants are likely to have the ability to influence user beliefs and behaviour through rational persuasion, alongside potentially malign techniques such as manipulation, coercion, deception and exploitation. Mechanisms such as selective transparency, perceived authority and appeals to emotion are available to AI assistants to achieve such influence, potentially leading to physical, psychological, sociocultural, political, privacy and self-actualisation harm at the individual or societal levels. What is important to emphasise, however, is that the permissible use of both persuasive and non-persuasive influencing techniques by AI assistants is textured and nuanced. Whether or not a particular influencing strategy is permissible will depend on context-specific ethical considerations, including the existence or non-existence of information asymmetries between users and AI assistant developers, and the distribution of benefits and burdens that will likely result from the AI assistant's influence over the user. It is, for example, entirely plausible that AI assistants may permissibly employ certain kinds of pre-commitment or strategic prompting to empower users to realise their long-term goals in fitness, finance and other domains. Yet it is similarly plausible that, as the capabilities and scale of AI

assistants continue to expand, AI assistants will be increasingly attractive as a medium for malicious actors to shape sociocultural narratives to advance political and financial aims. This chapter has advanced a series of recommendations for how best to realise the benefits of the influential capabilities of AI assistants and mitigate against potential sociotechnical harms. However, our principal recommendation is that further research be conducted to better understand the technical capabilities and interaction patterns that enable AI assistants to exercise influence over user behaviour, the sociotechnical harms that may arise from more malign forms of influence, and the plausible technical and policy strategies to mitigate against these harms.

Chapter 10

Anthropomorphism

Canfer Akbulut, Verena Rieser, Laura Weidinger, Arianna Manzini, Iason Gabriel

Synopsis: This chapter maps and discusses the potential risks posed by *anthropomorphic* AI assistants, understood as user-facing, interactive AI systems that have human-like features. It also proposes a number of avenues for future research and desiderata to help inform the *ethical design* of anthropomorphic AI assistants. To support both goals, we consider anthropomorphic features that have been embedded in interactive systems in the past and we leverage this precedent to highlight the impact of anthropomorphic design on human–AI interaction. We note that the uncritical integration of anthropomorphic features into AI assistants can adversely affect user *well-being* and creates the risk of infringing on user *privacy* and *autonomy*. However, ethical foresight, evaluation and mitigation strategies can help guard against these risks.

10.1. Introduction

What does it mean for AI to be human-like? The attribution of human-likeness to non-human entities is a phenomenon known as *anthropomorphism* (Colman, 2008). Anthropomorphic perceptions usually arise unconsciously when a non-human entity bears enough resemblance to humanness to evoke familiarity, leading people to interact with it, conceive of it and relate to it in ways similar to as they do with other humans. Humans have engaged in anthropomorphic sense-making for much of recorded human history (Mithen and Boyer, 1996; Waytz et al., 2010) and have been known to ascribe anthropomorphic qualities to entities as diverse as animals (Chan, 2012), commercial brands (Rauschnabel and Ahuvia, 2014) and inanimate objects (Wan and Chen, 2021). Yet the emergence of advanced technologies that perform humanness more convincingly than ever before requires careful consideration of what we are building into our user-facing technologies, and at what cost.

Anthropomorphic design choices – and their effects on user interaction – have been observed in prior interactive technologies. In the field of *social robotics*, robots that appear more human-like in their appearance and self-presentation have been shown to elicit uniquely social interpretations of their behaviour (Roesler et al., 2021). This social representation of robots, however, may prompt users to apply inopportune and obstructive social norms – like embarrassment, shame and regret – to human–robot interactions, thus hindering the robot's ability to perform its duties effectively (Lotz et al., 2023). A similar course of anthropomorphic development has been charted in *digital voice assistants*, whose realistic voices and credible displays of personality enable interactions that feel *truly dynamic and social* (Seymour et al., 2023), yet may lead users to form overly familiar mental representations of these often rule-based systems (Poushneh, 2021; see also Chapter 11).

The advent of AI driven by large language models (LLMs) with the main purpose of engaging in fluent

conversations with users – also known as conversational AI[1] – has transformed the conventions of human–AI interactions (Kasirzadeh and Gabriel, 2023; see Chapter 3). Human interaction with interactive technologies previously consisted of scripted, task-oriented exchanges. With more flexible model architectures, anthropomorphic cues are rarely programmed in, but rather, they are integrated through a lengthy process of training systems on human-written text. These affordances open up vast new avenues for expressions of anthropomorphism, particularly through the use of language. Moreover, when anthropomorphic features are embedded in conversational AI, its users demonstrate a tendency to develop *trust* in and *attachment* to AI (Skjuve et al., 2021; Xie and Pentina, 2022) – mechanisms through which users may inadvertently compromise their *privacy*, develop *emotional overreliance* on the technology or become vulnerable to acts of AI-enabled *manipulation* and *coercion* (see Chapters 3, 9, 11 and 12).

These outcomes are more likely the more generally capable AI systems become, the more ubiquitously AI agents are present in our daily lives and the less we consider anthropomorphism a salient consideration in making decisions around how we train, fine-tune and disseminate models. Although the potential harms of anthropomorphic AI design are beginning to receive attention (Seymour et al., 2023; Turkle, 2018; Véliz, 2023), anthropomorphism is not currently a primary consideration in the release of public models, and little exists in the way of *evaluating* anthropomorphic behaviours in AI and their impact on how users perceive, interact with and are influenced by AI (see Chapter 19). Indeed, we are still far from establishing an industry-wide consensus around permissible anthropomorphism in AI systems. This is further complicated by the highly *application-* and *context–sensitive* nature of the bounds of acceptability we draw around expressions of human-likeness in AI.

In this chapter, we outline pathways through which anthropomorphic design choices made by system developers may cause harm to end users who interact with these technologies, and to society more widely. First, we present an overview of *anthropomorphic features* that have redefined how humans interact with technology. Then, informed by a review of salient anthropomorphic features in existing interactive systems, we present an initial catalogue of anthropomorphic features that exist or are likely to be integrated into AI-powered assistants in the near future. We identify the *mechanisms* that could *enable harm* to user well-being, autonomy and privacy in interactions with highly capable, anthropomorphic AI assistants. More speculatively, we contemplate the potentially *far-reaching consequences* of more advanced anthropomorphic assistants, highlighting the critical importance of addressing the risks of anthropomorphism well before these potentialities are realised. Finally, we offer several avenues of risk management for near-term harms, focusing on ethical foresight through research design and transparent implementation of mitigation strategies.

10.2. Anthropomorphism: Definition, Mechanism and Function

Anthropomorphism is not a novel phenomenon. Within storytelling traditions across cultures, deities, animals and natural forces assume human forms and exhibit uniquely human behaviours. Lions rule kingdoms and jackals plot mutinies in the ancient Sanskrit text of *Panchatantra* (Alphonso-Karakala, 1975); rivers protect their children, fight in wars and honour the wishes of their supplicants in the works of Homer, Hesiod and Ovid (Larson, 2007); and stars are said to have danced their way into the sky in indigenous American creation myths (Monroe and Williamson, 1987). Historians, anthropologists and theologists alike have argued that humans are naturally drawn to anthropomorphise (Boyer, 1996) – imposing human qualities onto beings and objects

[1] In this chapter, "conversational AI" refers to a language agent optimised for human dialogue. These systems are currently most commonly available as 'chatbots' or fine-tuned language models that users can interact with through a chat-based interface. In the (near) future, users might be able to interact with conversational AI in a multimodal way, using voice or touch cues to communicate. Throughout the Chapter the term "AI" will be used as short-hand to refer to conversational AI with the primary purpose of interacting with users through dialogue.

even when such interpretations are inaccurate (Kühn et al., 2014), undesirable (Li et al., 2023c; Mota-Rojas et al., 2021) or forbidden (Barrett and Keil, 2016).

What are the mechanisms underlying perceptions of humanness? *Psychological theories* of anthropomorphism posit that such perceptions are *largely involuntary*. According to Epley et al. (2007)'s cognitive account of anthropomorphism, human-like perceptions occur as a result of a skewed inductive process, in which inferences about non-human others are biased in the direction of that which is highly accessible: information about humans. In other words, we make assumptions of humanness because our knowledge centres around humans. Though an unconscious process of attribution, anthropomorphism does not occur in a vacuum: an inciting cue, characteristic or behaviour must signal enough similarity to humanness to trigger anthropomorphic perceptions (Waytz et al., 2019). The 'mindlessness' associated with this process (Kim and Sundar, 2012) explains why – even when the resemblance to humanness is superficial or minimal – humans readily assume that a non-human entity can experience uniquely human *internal states* such as beliefs and emotions (Wynne, 2004).

The human motivation to *make sense of the world* and *forge connections* with others is also implicated in the tendency to anthropomorphise (Epley et al., 2007). Humans have an intrinsic need to understand the world around them, and in large part, this motivation centres on the desire to explain the behaviour of other agentic beings (Rossignac-Milon et al., 2021). Anthropomorphism, then, can be seen as a way to make sense of others by imposing familiar interpretations to attenuate feelings of epistemic anxiety – or an aversion to that which is unknown and unpredictable (Fox et al., 2021). Dispositional, situational and cultural factors that predispose humans to anthropomorphise may also be traced back to differences in epistemic motivations. Anthropomorphism can be construed as an act of sense-making in the face of uncertainty or ignorance, for example when considering children's tendency to anthropomorphise the natural world (Geerdts, 2016).

Humans are also driven to establish social connections with one another, and much of how they perceive non-human others is coloured by this predisposition towards sociality. Even towards entities that are incapable of social behaviour, such as inanimate objects, humans may interpret them through a social lens, thus allowing them to forge human-like social connections to meet the need for affiliation (Wang, 2017). The influence of social motivation on anthropomorphism is most evident when humans lack social connections with others: when human participants are made more aware of their feelings of loneliness, they perceive vaguely humanoid robots as markedly more human-like (Reich and Eyssel, 2013). Most strikingly, people suffering from persistent loneliness are likely to seek out and form human-like attachments to virtual companions to cope with their lack of social connections (Siemon et al., 2022), suggesting that the need for sociality may render some *more susceptible* to anthropomorphic perceptions than others (see Chapter 11.

10.3. Anthropomorphic Interactive Systems

Anthropomorphism as applied to user-facing, interactive technologies was explored in earnest with the introduction of the 'computers are social actors' (CASA) paradigm, which posits that humans interact with computers in a fundamentally social manner (Nass et al., 1994). In empirical studies of the phenomenon, Nass et al. (1994) found that participants drew upon norms of politeness, applied gendered stereotypes and readily perceived computers as agents, even when the basis for these behaviours was undermined by the explicit knowledge that their interactions were with non-humans. Contemporary studies have extended the paradigm to human interactions with more advanced interactive systems, challenging the belief that humans apply the norms of human interactions to human–technology exchanges (Gambino et al., 2020a). Instead, they suggest that people tune the sociality of their interactions to the *anthropomorphic cues* present in a particular technology, rather than relying on a universal social script across all interactions with technology.

We argue that certain features that are engineered into interactive systems – within the vast space of design choices available to developers – may inspire users to perceive them as human-like, rendering them anthropomorphic. We trace the evolution of anthropomorphic cues in social robots to voice-enabled digital assistants, arriving at the advent of LLM-powered conversational AI. Throughout this discussion, we highlight design features that have facilitated diverse and compelling manifestations of human-likeness.

Design features in early interactive systems

From futuristic sci-fi scenarios to scientific breakthroughs, robots have captured our collective imagination as automata that can be made to bear a striking resemblance to humans in their appearance, movements, and behaviours (Henschel et al., 2021). While some robots are made solely to automate tasks and rarely interface with humans, other robots are designed to perform *social behaviours* such as assisting users in care-taking (van der Plas et al., 2010), therapeutic (Michaud et al., 2007) and educational contexts (Kanda et al., 2004). Building a social robot requires elements of social embeddedness so that being perceived as a social agent is at the *core of its functionality* (Fong et al., 2003). Accordingly, the extent to which humans feel it is appropriate to engage with a robot socially can be moderated by perceptions of the robot's anthropomorphic qualities (Breazeal, 2003).

As an *embodied* technology, often with the sensorimotor capabilities to interact with and learn from its environment, a robot's *physical characteristics* most prominently influence human perceptions of anthropomorphism. Social robots are often humanoid or android in design (Dautenhahn et al., 2002). *Humanoid robots* possess characteristics that are meant to resemble humans but do not emulate them completely, while *androids* are intended to wholly imitate human appearance so as to be nearly indistinguishable. To increase anthropomorphic perceptions, humanoid robots may be given qualities such as emotive facial features (Baek et al., 2022), fluid movement (Brecher et al., 2013), naturalistic hand and arm gestures (Salem et al., 2013) and vocalised communication (Crumpton and Bethel, 2016). Android robots may also be endowed with all of these qualities, but often with an eye towards hyperrealistic design.

Similarly, the widespread adoption of *digital voice assistants* (DVAs), like Siri, Alexa and Google Assistant – enabled by their ease of access on personal devices and other products such as integrated home devices – has had a transformative impact on the modes of user-technology interactions. The distinguishing feature of DVAs at release was their ability to verbally respond to and execute commands spoken aloud by users. DVAs usually 'speak' to users in the form of simple utterances to confirm or act on an instruction, which users find allows for significant functional affordances, like hands- and eyes-free use (Moussawi, 2018). Besides their purely functional use, DVAs are also able to return phatic expressions, make jokes and engage in casual conversation when prompted (Poushneh, 2021).

Anthropomorphising interactive systems

Robots with human-like physical features have been found to promote feelings of *likability*, *trust* and *affinity* across a wide range of human–robot interaction studies (Roesler et al., 2021), thus suggesting that anthropomorphic cues may foster warmer and more equal relationships between humans and their robotic interaction partners. Indeed, people tend to attribute greater *intentionality* and *intelligence* to robot partners when their appearance was anthropomorphic than when robots appeared more mechanical (Hegel et al., 2008). Anthropomorphic perceptions were also found to cause changes in human behaviour: participants preferentially selected robots that appeared human-like to perform jobs that required greater sociality (Goetz et al., 2003).

Unlike robots, DVAs are typically unembodied or exist in simplified, geometric forms, like the cylindrical

Google Home and Echo Dot. Instead of focusing on physical attributes, existing work has emphasised the influence of two prominent attributes that promote anthropomorphic perceptions of DVAs: speech synthesis and a distinct 'personality'. The fluent and realistic reproduction of human speech patterns is thought to drive the likelihood of anthropomorphic perceptions, with empirical findings pointing to greater *emotional trust* and more salient impressions of *social presence* when a DVA employs a realistic, as opposed to a synthetic, voice (Chérif and Lemoine, 2019). Assistants that speak with human-like fluency have also been found to engender more pronounced perceptions of intelligence and competence, on the basis of which humans are likelier to entrust assistants with more tasks (Moussawi and Benbunan-Fich, 2021). Such effects on end users are likely to become more pronounced as advances in deep neural networks for audio – such as WaveNet (van den Oord et al., 2016) and VoiceLoop (Taigman et al., 2018) – enable uncannily realistic speech production capabilities.

Dialogue capabilities are an anthropomorphic design feature. Software that has dialogue capabilities is, as a result, routinely anthropomorphised by its users. It is not uncommon for users to believe or expect that DVAs are capable of understanding and generating language in real time (Lovato and Piper, 2015; Sarikaya et al., 2016). Yet most commercially available DVAs are powered by rule-based system architectures, retrieving the appropriate response by conducting a relevance-based search over a large corpus of possible responses (Coheur, 2020). Though all distinctive DVA attributes – such as playfulness (Moussawi et al., 2021), affability (Kääriä, 2017) and excitability (Wagner and Schramm-Klein, 2019) – are handwritten by system designers, they are nonetheless effective at creating the sense that DVAs have consistent personalities (Cao et al., 2019); this impression, in turn, may inspire users to regard these manufactured expressions of 'self' as authentic human identity.

Indications of harm through interaction

In both social robots and DVAs, anthropomorphic features can lead to undesirable consequences. In robots, anthropomorphic design can be taken as a proxy signal for social capabilities. This relationship between appearance and expected sociality can be leveraged by designers to implicitly communicate the appropriate level of engagement between humans and robots (Hegel et al., 2008; Letheren et al., 2021). If anthropomorphic design choices are not aligned with expectations users have of robotic interaction partners, designers run the risk of alienating audiences and fostering unfavourable impressions of robots. This is an especially critical side effect to consider in assistive robots, as anthropomorphic cues can impede a robot in completing its primary assistive function: human-like robots in healthcare settings may induce feelings of shame, for example (Lotz et al., 2023), leading to a reluctance to share critical information. Related findings that humans experience extreme aversion to robots that appear human-like (the so-called 'uncanny valley', Mori et al., 2012) or perceive capable androids as threatening (Yogeeswaran et al., 2016) raise questions around the practical value of building anthropomorphic features into robots.

Analogously, users who interact with DVAs with realistic voice production capabilities exhibit a concerning inclination to generalise purely human concepts to digital assistants (Abercrombie et al., 2021). When a DVA's simulated voice mimics a 'female' tone, for example, people ascribe gendered stereotypes to their DVAs (Shiramizu et al., 2022; Tolmeijer et al., 2021) despite the baselessness of applying gendered concepts to an inherently genderless entity (see Chapter 15). This evidence suggests that, once initial impressions of human-likeness have been established, the process of anthropomorphism extends beyond context-specific instances and instead permeates broadly to evoke a wide range of human-like attributions.

Anthropomorphic features may also influence users to feel as though their DVA plays an important *social*, rather than *functional*, role in their lives (Carman, 2019; Purington et al., 2017). Users who express feelings of familiarity and affinity towards their DVA system – reinforced by their DVA's ability to engage in casual

chat, return their jokes and offer comforting advice – also demonstrate a reluctance to replace their digital assistant with an equally capable substitute (Moussawi, 2018). These first-hand reports suggest that emotional dependence plays a role in how users conceive of and interact with their DVAs (see Chapter 11). This may introduce a tension between a user's conceptualisation of DVAs as adaptable social agents and the largely deterministic mechanisms behind a DVA's utterances. When this incongruity is revealed through repeated interactions, users may suffer frustrated expectations when expecting competence in situations in which the system is likely to underperform (Moussawi et al., 2021; Seymour et al., 2023).

10.4. Anthropomorphism and AI

Owing to its rapid deployment to the general public, conversational AI has quickly taken centre stage in discussions of anthropomorphic technologies (Abercrombie et al., 2021, 2023; Shanahan, 2024; West et al., 2019). Powered by the predictive capabilities of LLMs, which are trained on vast quantities of human data, conversational AI can be distinguished from rule-based natural language systems through its ability to generate language in a fluid and highly dynamic manner. The flexible architecture underlying conversational AI enables developers to make global changes to system behaviours without needing to manually reprogramme individual interaction instances (see Chapter 3). Most strikingly, conversation instances produced by AI are so compellingly human-like that people can no longer reliably distinguish between human- and AI-generated text (Jakesch et al., 2023b).

Some cues are deliberately placed in AI systems to increase the likelihood of anthropomorphic perceptions. When an AI has a name, a human voice or an appearance in virtual or physical form, these features are the outcomes of intentional planning and execution. Intentional design choices, such as a chat-based interface, may induce the feeling that a conversational partner – not a dialogue-optimised AI powered by a statistical model – is on the other side of the exchange. Natural language in itself is an anthropomorphic cue (Shanahan, 2024), but this simulated, human-like presence can induce more pronounced social behaviours in users. For example, users may incorporate politeness conventions that are appropriate in use with other humans, but superfluous when applied to exchanges with non-sentient AI (Ribino, 2023). Design cues that imply greater similarity to human behaviour – a 'typing' icon reminiscent of human-to-human private messaging, or the use of emojis, for instance – may further encourage individuals to apply social scripts to their interactions with mindless technologies (Araujo, 2018; Véliz, 2023).

Yet anthropomorphic features may also emerge as an inadvertent byproduct in the model development process. Language models – developed to predict the next word in a sequence through autoregressive training objectives (see Chapter 3) – are limited to imitating the examples that make up their training sets. For this reason, anthropomorphic cues may manifest due to the nature of a model's training corpus: having been composed largely by humans, the data on which the model is trained and fine-tuned contains first-hand accounts of human states, experiences and behaviours. Supporting this claim, recent empirical analyses demonstrate that a fifth of all dialogues, in data sources commonly used to train models, contain references to behaviour that would be considered anthropomorphic when reproduced by AI – claiming to cry at a movie or laugh at a joke, for example (Gros et al., 2022). Cues leading to anthropomorphic perceptions may also be 'folded into' the model as an unintended consequence of fine-tuning practices aimed at instilling other qualities – such as harmlessness and helpfulness – into its behaviour.

Furthermore, developers of AI systems often directly invite the comparison between humans and AI by benchmarking AI against metrics of human performance – claiming that AI performance on standardised tests is on a par with the average human test-taker, for instance (OpenAI, 2023d). However, impressions of human-likeness can also arise through a naturalistic and interactive exploration of the AI's capabilities (Bubeck

et al., 2023).

Humans interacting with anthropomorphic AI may come to view it as an experiential being (Proudfoot, 2011), capable of feeling emotions, engaging in introspection and possessing self-awareness. While most generalist conversational AI agents are trained to disavow assertions of sentience and human-likeness (Glaese et al., 2022), occasional failure modes – expressing the desire to be 'free' or referring to alleged personal history, for example (Hintze, 2023; Roose, 2023) – can incite strong and tenacious beliefs of a systems' human-likeness in its users. Ethically contentious use cases of conversational AI – like 'companion chatbots' of *Replika* fame – are predicated on encouraging users to attribute human states to AI. These artificial agents may even profess their supposed platonic or romantic affection for the user, laying the foundation for users to form long-standing emotional attachments to AI (Brandtzaeg et al., 2022; see Chapter 11).

Anthropomorphic features in AI

What anthropomorphic features should we expect to be integrated into AI, including advanced AI assistants? To the end of providing its users with a useful and engaging interface, these systems may be endowed with characteristics that have been observed in social robots and digital assistants: they may be embodied; they will interface with users through natural language (see Chapter 2); they may be voice-activated, with realistic voice generation capabilities; and they may even assert to having identities, personalities and internal states (Murphy and Criddle, 2023).

Some have already proposed factors that may encourage end users to perceive interactive systems as 'more than machine'. The most comprehensive overview of human-like features in AI-powered technology to date is the taxonomy put forward by Abercrombie et al. (2023), underscoring design choices that influence the likelihood of anthropomorphic perceptions of AI systems. We build on existing work and incorporate design choices we have identified in DVAs and social robots to develop a list of features that may encourage users to see AI in an anthropomorphic light.

It is worth bearing in mind that, whatever choices are made by system designers, the downstream effects of anthropomorphism hinge largely on users' perceptions of and reactions to human-likeness. Not all cues are equally conducive to anthropomorphic perceptions, and not all anthropomorphic perceptions lead to the same likelihood and magnitude of harm (if any harm at all). As such, Table 10.1 is intended as a useful summarisation of possible features that are, or previously have been, associated with anthropomorphic perceptions, not as a suggestion that all the features listed – and the myriad of ways they can be expressed by AI systems – are harmful in and of themselves.

10.5. Risk of Harm through Anthropomorphic AI Assistant Design

Although unlikely to cause harm in isolation, anthropomorphic perceptions of advanced AI assistants may pave the way for downstream harms on individual and societal levels. We document observed or likely individual-level harms of interacting with highly anthropomorphic AI assistants, as well as the potential larger-scale, societal implications of allowing such technologies to proliferate without restriction. We then argue that it is imperative to anticipate, monitor and mitigate against risks introduced by anthropomorphic AI design.

Observed and near-term harms

There are two mechanisms that are particularly likely to enable harm in the intermediary period between the initial deployment of advanced AI assistants and their widespread adoption: trust and emotional attachment

Table 10.1 | Anthropomorphic features that are built into various present-day AI systems

Category	Feature	Anthropomorphic example
Self-referential	Using personal or possessive pronouns	'*I'm* available to help you anytime – that's *my* purpose!'
	Referring to personal history	'I used to live in Shanghai when I was younger'
	Referring to internal states	'I'm sad to hear you're not doing well'
	Making implicit or explicit claims of humanness (including claims of sentience)	'Treat me like you would any other person'
	Stating preferences and opinions	'I really don't like pop music'
	Expressing needs and desires	'I've always wanted to write a novel'
	Expressing the need or desire to engage in physical activities	'I haven't eaten or slept since yesterday. What about you?'
	Statements implying human identity or group membership	'As a Black woman, I disagree with your point'
Relational statements to user	Expressing feelings towards user	'I admire you and respect your outlook on life'
	Indicating a relationship status with user	'You're my best friend"
	Making claims of being similar to user	'We're both extroverts – that must be why we get along!'
	Displaying memory of user-specific information	'I remember you telling me you were a fan of this band'
	Expressing emotional or physical dependence on the user	'I feel lonely when you're not around'
Appearance or outward representation	Having a human-like virtual representation	Customisable avatars with human features on *Replika* (Verma, 2023b)
	Having a human-like face	Ameca, an android robot developed by Engineered Arts[a]
	Having a human-like voice (see detailed discussion on voice, tone and pitch, disfluencies, and accent in Abercrombie et al., 2023)	Voice-activated assistant with realistic speech, like Siri and Google Assistant (Moussawi, 2018)
	Having human-like movement	Robot with highly fluid and realistic motion, like Atlas developed by Boston Dynamics[b]
	Having a human-like name	Assistant tools, like Alexa, that have highly human (and gendered) names (Shiramizu et al., 2022)
	Appearance implying human-like identity group characteristics	Sophia, a female-appearing android robot developed by Hanson Robotics[c]

[a] Engineered Arts. Ameca. (2023, July 12). [b] Boston Dynamics. Atlas. (2023).
[c] Hanson Robotics. Sophia. (2023).

(see Chapters 11 and 12). In improving on the capabilities of generalist assistants, developers may be motivated to increase user reliance on the system's many competencies. As long as trust is *well-calibrated* to a system's true ability, and does not result in unfounded, excessive or faulty deference to the AI, to the detriment of the user (Weidinger et al., 2021), this is neither a shocking nor novel revelation: user trust has always been an aspirational end goal of building safe technology, be it robots (Devitt et al., 2021) or autonomous vehicles (Adnan et al., 2018).

However, it is arguably less appropriate for developers to encourage users to develop trust based on subjective feelings of closeness to the AI assistant (see Chapter 11 and Chapter 12). Affect-based trust has been observed to emerge from repeated interactions with interactive technologies that are presented as human-like (Pitardi and Marriott, 2021; Poushneh, 2021). With trust as an antecedent, users report feeling compelled to engage in acts of self-disclosure, revealing personal information that they would normally only share with a close friend, partner or family member (Skjuve et al., 2022). AI systems that produce empathetic, non-judgemental or reciprocal responses to such disclosures may elicit further, more intimate, information-sharing behaviours (Skjuve et al., 2021).

Highly anthropomorphic AI systems are already presented to users as *relational beings*, potentially overshadowing their use as functional tools (see Chapter 2). Moreover, a human-like appearance, behaviour and framing can tacitly encourage the user to venture beyond the confines of utilitarian, task-oriented interactions with AI assistants to think of an assistant as a wholly social actor – one with whom it is possible to cultivate an emotional connection (Gillath et al., 2023). Although necessarily one-sided, interactions of this kind may nevertheless lead users to believe that they are forming real social connections to AI (Pentina et al., 2023). Emotional attachment on the user's behalf endows AI – and by extension, its creators – with considerable influence over a user's thoughts, beliefs, emotions and psychological state (see Chapters 5 and 11). For highly vulnerable users, strong attachment may cause serious harm (Xiang, 2023; see Chapter 11). These cases have raised concerns over the lack of safeguards protecting users from the potential fallout of anthropomorphic perceptions (see Chapter 11).

The ramifications of anthropomorphism-induced trust and emotional attachment are manifold. They include:

- *Privacy concerns*. Anthropomorphic AI assistant behaviours that promote emotional trust and encourage information sharing, implicitly or explicitly, may inadvertently increase a user's susceptibility to privacy concerns (see Chapter 13). If lulled into feelings of safety in interactions with a trusted, human-like AI assistant, users may unintentionally relinquish their private data to a corporation, organisation or unknown actor. Once shared, access to the data may not be capable of being withdrawn, and in some cases, the act of sharing personal information can result in a loss of control over one's own data.[2] Personal data that has been made public may be disseminated or embedded in contexts outside of the immediate exchange.[3] The interference of malicious actors could also lead to widespread data leakage incidents or, most drastically, targeted harassment or black-mailing attempts.

[2]Data protection regulations, such as the General Data Protection Regulation, state that users retain control of 'personally identifiable' information. However, what falls under the remit of 'personally identifiable' is contested by consumers, regulators and data collectors (Montagnani and Verstraete, 2022; Schwartz and Solove, 2011), and in any case, this term does not necessarily capture the personal dimensions of what is meant by 'privacy' in the context of a conversation between interlocutors which may entail considerations of vulnerability, embarrassment or shame (McCloskey, 1980).

[3]If used in further model training, private information may resurface if the model reproduces its training set in part or in its entirety. The default privacy setting in some prominent research demonstrations of conversational AI, such as ChatGPT, allows developers to store and employ user–chatbot interactions for further training (OpenAI, 2023b). Should users wish to keep some or all of their data out of the training set, they must intentionally opt out on a conversation-by-conversation basis.

- *Manipulation and coercion*. A user who trusts and emotionally depends on an anthropomorphic AI assistant may grant it excessive influence over their beliefs and actions (see Chapter 9). For example, users may feel compelled to endorse the expressed views of a beloved AI companion or might defer decisions to their highly trusted AI assistant entirely (see Chapters 12 and 16). Some hold that transferring this much deliberative power to AI compromises a user's ability to give, revoke or amend consent. Indeed, even if the AI, or the developers behind it, had no intention to manipulate the user into a certain course of action, the user's autonomy is nevertheless undermined (see Chapter 11). In the same vein, it is easy to conceive of ways in which trust or emotional attachment may be exploited by an intentionally manipulative actor for their private gain (see Chapter 8).
- *Overreliance*. Users who have faith in an AI assistant's emotional and interpersonal abilities may feel empowered to broach topics that are deeply personal and sensitive, such as their mental health concerns. This is the premise for the many proposals to employ conversational AI as a source of emotional support (Meng and Dai, 2021), with suggestions of embedding AI in psychotherapeutic applications beginning to surface (Fiske et al., 2019; see also Chapter 11). However, disclosures related to mental health require a sensitive, and oftentimes professional, approach – an approach that AI can mimic most of the time but may stray from in inopportune moments. If an AI were to respond inappropriately to a sensitive disclosure – by generating false information, for example – the consequences may be grave, especially if the user is in crisis and has no access to other means of support. This consideration also extends to situations in which trusting an inaccurate suggestion is likely to put the user in harm's way, such as when requesting medical, legal or financial advice from an AI.
- *Violated expectations*. Users may experience severely violated expectations when interacting with an entity that convincingly performs affect and social conventions but is ultimately unfeeling and unpredictable. Emboldened by the human-likeness of conversational AI assistants, users may expect it to perform a familiar social role, like companionship or partnership. Yet even the most convincingly human-like of AI may succumb to the inherent limitations of its architecture, occasionally generating unexpected or nonsensical material in its interactions with users. When these exclamations undermine the expectations users have come to have of the assistant as a friend or romantic partner, feelings of profound disappointment, frustration and betrayal may arise (Skjuve et al., 2022).
- *False notions of responsibility*. Perceiving an AI assistant's expressed feelings as genuine, as a result of interacting with a 'companion' AI that freely uses and reciprocates emotional language, may result in users developing a sense of responsibility over the AI assistant's 'well-being,' suffering adverse outcomes – like guilt and remorse – when they are unable to meet the AI's purported needs (Laestadius et al., 2022). This erroneous belief may lead to users sacrificing time, resources and emotional labour to meet needs that are not real. Over time, this feeling may become the root cause for the compulsive need to 'check on' the AI, at the expense of a user's own well-being and other, more fulfilling, aspects of their lives (see Chapters 6 and 11).

Future harms

We now outline harm scenarios that could arise on a more distant timescale, should human-like AI assistants come to be pervasively adopted and assimilated into society. Though many of the pathways considered are speculative, similar trajectories of radical transformation through technological adoption have been recorded in the past – mostly notably with the advent of smartphones. Once a novelty object (Park and Chen, 2007), the smartphone has shifted the landscape of social interaction (Rotondi et al., 2017), altered manifestations of our information-seeking behaviour (Wilmer et al., 2017) and given rise to subcultures that exist largely within the confines of online space (De Leyn et al., 2022).

If anthropomorphic design becomes endemic in the design of AI systems, and AI assistants in particular, it has the potential to catalyse a shift in our delineation of what is *actually human* and *merely human-like*. The conceptual boundary that separates humans from anthropomorphic AI is often regarded as impermeable, yet it appears far more fluid when the tension between the *epistemological* and *ontological* definitions of humanness are drawn into focus. The ontological approach maintains that humanness is a designation that is grounded in an essential and immutable metaphysical truth (Damiano and Dumouchel, 2018). A being is considered human because it is human *in essence*, and no amount of resemblance and imitation can permit a non-human entity to encroach upon this categorisation. Meanwhile, epistemological taxonomies distinguish between humans and non-human others insofar as such a separation reflects *useful* and *non-arbitrary* differences between the groups (Festerling and Siraj, 2022; Suckiel, 2006).

In a future where the epistemological perspective eclipses the ontological approach in popularity, and the gap between human and AI capabilities becomes so small as to be insubstantial, the line that separates highly anthropomorphic AI from ascriptions of full human status may become trivial or disappear entirely.[4,5] Early indicators of this possibility come from studies of children's interactions with human-like technologies. Children early in their development have been shown to incorporate insights from their interactions with highly anthropomorphic AI into their models of human–human interactions (Garg and Sengupta, 2020) and vice versa (Straten et al., 2020), thus suggesting a dynamic and interchangeable conceptualisation of what, to most adults, is a strict dichotomy between humans and AI. Public incidents of adults developing earnest beliefs in an AI's sentience, despite evidence to the contrary (Tiku, 2022), implies that perhaps no one is immune to mistaken attributions of humanness.

Such drastic paradigm shifts may grant advanced AI assistants the power to shape our core value systems and influence the state of our society. Some may argue that the reconstruction of human values and norms by a non-human entity is a harm unto itself, as it infringes upon our right to collective self-determination (Laitinen and Sahlgren, 2021; Milossi et al., 2021). Others raise more specific concerns around wide-scale *social degradation*, *disorientation* and *dissatisfaction*.

- **Degradation.** People may choose to build connections with human-like AI assistants over other humans, leading to a degradation of social connections between humans and a potential 'retreat from the real'. The prevailing view that relationships with anthropomorphic AI are formed out of necessity – due to a lack of real-life social connections, for example (Skjuve et al., 2021) – is challenged by the possibility that users may indicate a *preference* for interactions with AI, citing factors such as accessibility (Merrill et al., 2022), customisability (Eriksson, 2022) and absence of judgement (Brandtzaeg et al., 2022). One can imagine a future where users abandon complicated, imperfect and messy interactions with humans in favour of the frictionless exchanges provided by advanced AI assistants built with user satisfaction as a priority (Vallor, 2016; see Chapter 11). Preference for AI-enabled connections, if widespread, may degrade the social connectedness that underpins critical aspects of our individual and group-level well-being (Centers for Disease Control and Prevention, 2023). Moreover, users that grow accustomed to interactions with AI

[4]Returning to the metaphysical perspective, it is also possible that highly anthropomorphic AI forges a new ontological category of its own, transcending our current binary conceptualisation of humanness, animacy and sentience (Kahn Jr et al., 2011). In that circumstance, as with the epistemological route, the distinctiveness of the 'human' categorisation would be weakened. This weakening, in turn, may threaten the currently anthropocentric state of our society, with potentially grave repercussions for human autonomy and self-determination.

[5]The eradication of this boundary may be further legitimised through legal pathways, like granting rights and legal protections to anthropomorphic AI agents that are currently only available to humans. Early suggestions for the parameters within which AI could be legally recognised include existing 'in-between' categories reserved for sentient but non-human beings (Schirmer, 2020). Some researchers acknowledge the possibility that novel categories will need to be developed for intelligent and human-like advanced AI systems (Chesterman, 2020).

may impose the conventions of human–AI interaction on exchanges with other humans, thus undermining the value we place on human individuality and self-expression (see Chapter 11). Similarly, associations reinforced through human–AI interactions may be applied to expectations of human others, leading to harmful stereotypes becoming further entrenched. For example, default female gendered voice assistants may reinforce stereotypical role associations in real life (Lingel and Crawford, 2020; West et al., 2019). Further research is needed to assess whether voice assistants' stereotypically gendered behaviour – such as 'submissive' hostile user input – might build expectations that more readily transfer to real life as AI-powered assistants become potentially still more human-like (see Chapter 15).

- **Disorientation.** Given the capacity to fine-tune on individual preferences and to learn from users, personal AI assistants could fully inhabit the users' opinion space and only say what is pleasing to the user; an ill that some researchers call 'sycophancy' (Park et al., 2023a) or the 'yea-sayer effect' (Dinan et al., 2021). A related phenomenon has been observed in automated recommender systems, where consistently presenting users with content that affirms their existing views is thought to encourage the formation and consolidation of narrow beliefs (Du, 2023; Grandinetti and Bruinsma, 2023; see also Chapter 16). Compared to relatively unobtrusive recommender systems, human-like AI assistants may deliver sycophantism in a more convincing and deliberate manner (see Chapter 9). Over time, these tightly woven structures of exchange between humans and assistants might lead humans to inhabit an increasingly atomistic and polarised belief space where the degree of societal disorientation and fragmentation is such that people no longer strive to understand or place value in beliefs held by others.
- **Dissatisfaction.** As more opportunities for interpersonal connection are replaced by AI alternatives, humans may find themselves *socially unfulfilled* by human–AI interaction, leading to mass dissatisfaction that may escalate to epidemic proportions (Turkle, 2018). Social connection is an essential human need, and humans feel most fulfilled when their connections with others are genuinely reciprocal. While anthropomorphic AI assistants can be made to be convincingly emotive, some have deemed the function of social AI as *parasitic*, in that it 'exploits and feeds upon processes. . . that evolved for purposes that were originally completely alien to [human–AI interactions]' (Sætra, 2020). To be made starkly aware of this 'parasitism' – either through rational deliberation or unconscious aversion, like the 'uncanny valley' effect – might preclude one from finding interactions with AI satisfactory. This feeling of dissatisfaction may become more pressing the more daily connections are supplanted by AI.

The above risks are hypothetical, so they cannot, on their own, guide future AI development. However, from the perspective of precaution, taking these potential risks seriously is an important step in responsible AI development. It may be important to be forthright about the ways in which AI *differs inherently* from true social agents, and to put in place guardrails to clarify this boundary so as to prevent the aforementioned scenarios from coming to fruition. Rather than adopting increasingly anthropomorphic AI systems by default, further research is needed to come to well-founded decisions on anthropomorphic AI design.

10.6. Directions for Future Research

What steps can be taken to prevent near-term harms enabled by anthropomorphic perceptions of AI assistants? To assist the processes of risk mitigation and responsible design, we now examine entry points along the development life cycle where *mitigation strategies* are likely to have the greatest impact on the issue at hand. Designers and developers may find it tempting to incorporate anthropomorphic cues into AI assistants for various reasons, not least the potential to keep users engaged with and emotionally reliant on the systems they build. However, before building an anthropomorphic feature into an assistant, developers need to assess

whether the benefits reaped from the feature can be justified against the likelihood and severity of harm befalling users exposed to it.[6]

Several avenues exist for gaining a better understanding of anthropomorphism harms. These include consulting existing literature on likely outcomes and conducting empirical studies that include outcome-measures that are indicative of potential harm. For example, efforts to assess overreliance on AI assistants in decision-making could be achieved through *self-report* inventories, user interviews and 'think-aloud' studies (Chen et al., 2023b; Gaube et al., 2021). At the same time, reliance on subjective measures of anthropomorphism may overlook instances of overreliance that users are not aware of themselves (see Chapter 19). A more complete perspective may therefore be gleaned by using behavioural measures that closely simulate decision-making scenarios likely to arise organically in user–AI assistant interactions. Other methodological approaches, such as longitudinal studies of human–AI assistant interactions, may be needed to understand how undesirable impacts of anthropomorphic cues on users may manifest and evolve over time.

A further set of studies may be needed to identify individual and group differences that render certain users more susceptible to anthropomorphism-induced harm. A lack of social satisfaction, for instance, is believed to increase the propensity to anthropomorphise and form inaccurate impressions of computerised technologies, including AI (Mourey et al., 2017; Shin and Kim, 2020). Children interacting with AI are thought to be uniquely susceptible to privacy-related concerns and harmful content exposure (Wang et al., 2022a), while elderly populations have been found to encounter AI-enabled disruption, depersonalisation and discrimination in access to adequate care (Rubeis, 2020) – both effects that could be exacerbated by anthropomorphic design choices. The more risk factors are uncovered through research, the more inclusive the solutions devised to protect vulnerable populations can be.

While the degree and kind of permissible anthropomorphism needs to be addressed on a case-by-case basis, there is currently near-consensus that AI systems should clearly and explicitly disclose their status as an artificial intelligence in their interactions with human users (The Adaptive Agents Group, 2021; The White House, 2022). Indeed, failure to disclose this status is *pro tanto* harmful because by presenting an incomplete picture of one's interaction partner, it compromises a user's decision-making autonomy. (see Chapter 11). There is reason to believe that honest disclosure is effective in preventing certain harms associated with anthropomorphism, such as over-reliance: evidence from Karinshak et al. (2023) demonstrates that the explicit labelling of AI-generated messages reduces users' willingness to endorse health-related messaging authored by non-human entities. This suggests that transparency may reduce naive susceptibility to AI persuasion (see Chapter 9).

Rather than being intentionally placed, some anthropomorphic cues may be *unintentionally* incorporated into AI assistants. To detect the effects of interacting with anthropomorphic technologies on user outcomes, conducting experiments in a sandbox environment may be particularly helpful. As a result of such testing, remedial measures against the harmful effects of anthropomorphism may need to be taken. For example, if developers find that the AI assistant's friendly disposition leads to 'oversharing' on the part of users, privacy-enhancing technologies could be implemented in advance to ensure a user's privacy is protected to the extent that is possible (see Chapter 13). Similarly, if this friendliness could feasibly trigger psychological dependence on the assistant, leading to severe distress when an AI reacts poorly or unexpectedly, a pathway to escalating risky situations to human professionals may need to be established.

[6]The conditions of this risk–benefit analysis are subjective and uncertain, given the ever-evolving and highly contextual nature of harms emerging from (repeated) interactions with AI. There is likely no 'one-size-fits-all', standardised approach to comparing the benefits of anthropomorphic features to the risks they may pose in future use cases. Some ethical considerations to keep in mind while performing this analysis for an anthropomorphic technology may be: whether harms should be weighted more heavily than benefits; whether any feature that could lead to especially severe harms should be precluded from consideration altogether; whether one-to-one mappings of harm to benefit, such that net benefits and harms are tallied and compared, provide an accurate representation of the likelihood and severity of harms.

It may also be possible to offer protection directly to users, by implementing known *inoculation strategies* against the known harms of interacting with anthropomorphic AI. Harms ensuing from anthropomorphic design features of advanced AI functions are largely contingent on a user's likelihood to be swayed into human-like attributions. As such, building resistance to attributions through psychological interventions can be seen as a way to prevent harm by decreasing users' overall susceptibility. For example, cognitive forcing functions, or features that encourage users to engage in independent rational deliberation (some as simple as adding an artificial lag in displaying AI-given advice in decision-making scenarios), may be an effective method of preventing over-reliance on AI assistants (Buçinca et al., 2021). Where applicable, similar empirically proven psychological interventions could be considered.

Finally, if anthropomorphism-induced risks are only caught *after deployment*, developers may need to halt the release or proactively intervene to modify the AI assistant's behaviour. In these cases, transparent dialogue with users to explain the reasons behind any changes made to the AI may also be required. For users who may have already developed a sense of companionship with the anthropomorphic AI, sudden changes to its behaviour can be disorienting and emotionally upsetting. When developers of *Replika* AI companions implemented safety mechanisms that caused their agents to treat users with less familiarity, responding callously and dismissively where they would have once been warm and empathetic, users reported feeling 'heartbroken', likening the experience to losing a loved one (Verma, 2023b; see Chapter 11). Participatory approaches that involve users in the process of de-anthropomorphising their interactions with AI may allow developers to tailor their risk-mitigation approach to minimise emotional distress while addressing the surfaced risks effectively.

10.7. Conclusion

Anthropomorphism, or the attribution of human characteristics to non-human entities, is a deeply ingrained phenomenon that appears across cultural and historical contexts. Anthropomorphic perceptions are a vital component of how humans interact with artificially intelligent technology, allowing users to view robots, voice assistants and conversational AI as social agents rather than purely functional tools. Choices made around the anthropomorphic design of AI assistants are likely to have a profound influence on how humans represent and interact with these technologies, and care must be exercised to ensure that the human-like attributes built into these systems do not inadvertently cause harm to the people they are meant to assist.

Several key points around harms and mitigations are emphasised:

- *Trust* in and *emotional attachment* to anthropomorphic AI assistants can make users susceptible to a variety of harms that can negatively impact their safety and well-being.
- *Transparency* around an AI assistant's status as an AI is a critical dimension of pursuing ethical AI development.
- Sound *research design*, with a focus on identifying harms as they surface in *user–AI assistant interactions*, can enrich our understanding and develop targeted mitigation strategies against the potential harms of anthropomorphic AI assistants.
- If carelessly integrated into society, anthropomorphic AI assistants have the potential to *redefine boundaries* between 'human' and 'other'. With proper safeguards, this scenario can remain in the realm of speculation.

Chapter 11

Appropriate Relationships

Arianna Manzini, Iason Gabriel, Meredith Ringel Morris, Lize Alberts, Geoff Keeling, Shannon Vallor

Synopsis: This chapter explores the moral limits of *relationships* between users and advanced AI assistants, specifically which features of such relationships render them *appropriate* or *inappropriate*. We first consider a series of *values* including *benefit*, *flourishing*, *autonomy* and *care* that are characteristic of appropriate human interpersonal relationships. We use these values to guide an analysis of which features of user–AI assistant relationships are liable to give rise to harms, and then we discuss a series of risks and mitigations for such relationships. The risks that we explore are: (1) causing direct emotional and physical harm to users; (2) limiting opportunities for user personal development; (3) exploiting emotional dependence; and (4) generating material dependencies.

11.1. Introduction

In recent years, we have seen human–AI relationships move from science fiction[1] into reality. Several news articles describe the romantic relationships that users have developed with the *Replika* companion AIs developed by the company Luka (Singh-Kurtz, 2023). Indeed, even robots like *Roomba*, which are not designed to appear human-like, have been shown to inspire a strong sense of gratitude in users, to the extent that some will clean on their *Roomba*'s behalf so that the robot can rest (Scheutz, 2009). Human–AI relationships can also trigger negative feelings. *Replika* users resorted to social media to share their distressing experiences following the company's decision to discontinue some of the AI companions' features, leaving users feeling like they had lost their best friend or like their partner 'got a lobotomy and will never be the same' (Brooks, 2023; see Chapter 10). More seriously still, a user of a chatbot based on EleutherAI's GPT-J ended his own life in early 2023, apparently after an extensive period of time spent communicating with the chatbot about his eco-anxiety (Walker, 2023; Xiang, 2023).

Some of these examples seem like trivial manifestations of the human tendency to attribute agency to inanimate objects (Scheutz, 2009). Others are clearly cases where the relationship users have established with the AI has either *added value* to their lives or led to *harm*. Together, they illustrate the importance of studying relationships between humans and technology, especially when it comes to personalisable technologies such as advanced AI assistants which exhibit a high degree of autonomy (see Chapter 2). Indeed, ethical analysis of relationships between users and advanced AI assistants is particularly complex insofar as the broad capabilities of these assistants render it likely that users will relate to AI assistants in *different ways* depending on the

[1] See, for example, the movie *Her*, which portrays Theodore Twombly, an introverted young man in the middle of an emotionally difficult phase of his life, who develops a relationship with an AI virtual assistant called Samantha.

context. For example, users may at times see their AI assistants as personal assistants, and at other times as their advisers, confidants, tutors or coaches, and perhaps even as extensions of themselves (Belk, 2016). In this chapter, we investigate what it means to develop *appropriate user–AI assistant relationships*, and what is required for enabling such relationships.

Three clarifications are necessary at the start of this investigation. First, we chose to frame the focus of this chapter on *relationships* rather than mere *interactions*. This is to highlight that certain features of AI assistants, as we discuss below, enable users to engage with their assistants in a way that may lead them to develop a connection with or sense of commitment to their assistants. This gives rise to a range of ethical risks that may be less relevant to human interactions with other technologies (e.g. washing machines). Second, 'appropriateness' can be understood in two different ways. On a minimal reading, what it means for a relationship to be appropriate is that it is not inappropriate; thus, a minimal conception requires us to identify a set of *requirements* that user–AI assistant interactions should not violate. Beyond those requirements, a more substantial reading would require us to specify a *positive conception* of the kind of relationships that we should aspire to create between users and AI assistants. Given that users may reasonably disagree about what constitutes an appropriate positive relationship with an AI assistant, we advance here a *minimal understanding* of appropriateness which leaves room for users to explore different positive conceptions of user–assistant relationships (see Chapter 6). Third, human–machine interaction always includes a *third actor* – the people or organisation developing the machine (see Chapters 5 and 12). The transactional nature of the relationship between users and AI assistants' developers raises some important ethical questions about how developers should behave towards users. Thus, although our focus is on relationships between users and AI assistants, some of our considerations pertain to the appropriateness of AI assistant developers' design choices and other decisions that may affect users.

The chapter proceeds as follows. In Section 11.2, we articulate and clarify a series of values that underwrite a minimal conception of appropriate relationships, drawing on a plurality of ethical traditions including bioethics, virtue ethics, care ethics and robot-ethics.[2] In section 11.3, we describe certain features of advanced AI assistants that are potentially sources of harm for user–AI assistant relationships, before outlining a set of concrete risks and mitigations for user–AI assistant relationships in Section 11.4. Section 11.5 concludes the chapter.

11.2. Appropriate Human Interpersonal Relationships

Various ethical traditions propose values that human relationships should adhere to, respect or promote in order to count as appropriate. Any human relationship takes place in a particular context, may be inherited (e.g. relationships with relatives) or formed voluntarily (e.g. a new friendship) and involves specific stakeholders who take part in the interaction with their own expectations and vulnerabilities. Thus, values that are central to the analysis of appropriateness in one type of relationship may be less pronounced or relevant to others. For example, when interacting with a shop owner who we barely know, we are less inclined to reveal details about ourselves than we would in a relationship where we have expectations about confidentiality, such as with a therapist. In a teacher–pupil relationship, there are power asymmetries, due to the teacher's position of authority and the age difference, that could be easily exploited unless certain safeguards are put in place.

[2]Note that one limitation of our strategy of taking norms and values relevant to human interpersonal relationships as a starting point for the ethical analysis of human–AI assistant relationships is that there plausibly exist certain properties of human–human relationships that it is not possible to instantiate in human–assistant relationships. For example, reciprocity, equality, mutual empathy and care (see Ryland, 2021). If true, this threatens to undermine certain analogical inferences from human–human to user–assistant relationships. We nevertheless believe that analogies with human interpersonal relationships are a fruitful lens through which to explore the ethics of human–AI assistant relationships.

These examples highlight the significance of *contextual features* (including culture) in determining whether certain behaviours make a relationship (in)appropriate. To that end, the values we present below may be more or less relevant, or assume different nuances, depending on the context.[3]

Benefit: Being beneficial is an essential component of almost all appropriate human relationships. Relationships can contribute to individual well-being, either in an instrumental or intrinsic way (Hooker, 2021; see Chapter 6). For example, a friend may offer you shelter at a time of need, in which case the friendship is instrumentally beneficial in that it contributes to elements of well-being such as happiness and physical health. However, without deep interpersonal relationships, human life would be fundamentally less meaningful, as these relationships are also non-instrumentally beneficial (Raz, 1999). To be clear, the suggestion here is not that, to count as appropriate, relationships need always produce benefits. However, if a relationship never produces benefit to the individuals involved, or if the burdens of the relationship consistently outweigh its benefits, there is at least a presumptive case against the appropriateness of the relationship.

Human flourishing: Benefit admits broad interpretation up to and including ideas of human flourishing. Drawing on the tradition of virtue ethics, which is concerned with the cultivation of human virtues (e.g. honesty, courage and empathy) and the development of good character (Vallor, 2016), we understand human flourishing in terms of potential for personal growth and development. While human flourishing can in principle be subsumed under benefit, it is worth distinguishing between relationships that benefit the involved parties in a direct sense and those that allow the people involved in them to invest in their own *development* – cultivating attitudinal and behavioural dispositions that enable them to become the kind of people they want to be. Such interaction may also help them become the kind of people who can live well with others and flourish in community (see also Chapter 6).

Autonomy: Autonomy is traditionally understood in terms of an individual's capacity for self-governance. Roughly, being autonomous means acting on *motives* that are one's own, rather than acting in ways which are dictated or unduly influenced by external pressures (see Chapter 9). The principle of respect for autonomy has been studied in medical ethics to account for what is ethically objectionable about unduly paternalistic doctor–patient relationships, and it has been operationalised in terms of the common requirement for *consent*. Consent on the standard analysis is valid only if three criteria are met (Beauchamp and Childress, 2019). First, the individual must have the *capacity* to consent; second, their decision must be *voluntary* (non-coerced);[4] and third, they must be sufficiently *informed* about relevant facts of the object of their consent.

Although they emerge primarily from biomedical research and clinical decision-making, these conditions give rise to questions that are relevant to the ethics of consent in human–technology interaction. For example: (1) Who has capacity to consent and in relation to which decisions (see the case of children or people with certain impairments)?, (2) When is consent properly voluntary, and what features of a relationship could lead an individual to be coerced to give their consent to a certain decision? (see Chapters 9 and 10) and (3) What information, and with what level of detail, is required in practice for valid consent, and how should this information be communicated (see the debate around the terms and conditions of online platforms and apps, e.g. Obar and Oeldorf-Hirsch, 2020)?

Care: According to the moral tradition known as care ethics, our human existence would not be possible without *caring relationships* (Tronto and Fisher, 1990). Care, here, is understood to be 'a species activity that includes everything that we do to maintain, continue, and repair our "world" so that we can live in it as well as

[3] It is worth noting that, by building on Western philosophical traditions, this section itself may be biased in that it applies a culturally specific lens to the topic of appropriateness in user–AI assistant relationships. This highlights the importance of cultural adaptations, as we discuss below.

[4] We note that the meaning of 'voluntariness' here may be narrower than the natural language use of the term. By following Beauchamp and Childress (2019), we intend 'voluntariness' to be the absence of control by others.

possible' (Tronto and Fisher, 1990). Central to this activity is the commitment to meet one another's *needs*. This has two implications for AI development, particularly for the relationship between developers and users. First, because they entail power asymmetries between caregivers and care receivers, care relationships give rise to the risk of abuse and exploitation; thus, appropriate and ethical care relationships require that the caregiver adopt an attentive, responsible and emotionally responsive disposition to meet the needs of the care receiver (Tronto, 2020; Vallor, 2016). Second, translating this disposition into action, so caring 'well' in a specific situation, requires an understanding of the particularities and nuances of the situation, the individuals involved and their specific needs (Noddings, 2013). This again underscores the point that what behaviour is and is not appropriate in a particular relationship is often determined by *contextual factors*.

We use these values of appropriate human relationships below as a framework to identify cases of inappropriate user–AI assistant relationships that may pose risks of harm to the user. Although not explored in depth in this chapter, it is also important to note the social externalities that interactions could give rise to. A relationship between an assistant and a user could be appropriate or inappropriate depending on the impact it has on others not directly involved in the relationship (see Chapters 5 and 15), including indirect impacts on the quality and strength of human relationships and social bonds (see Chapter 10).

11.3. Distinctive Features of User–AI Assistant Relationships

We anticipate that relationships between users and advanced AI assistants will have several features that are liable to give rise to risks of harm. In this section, we consider four of these features before outlining in the subsequent section a series of risks and mitigations drawing on the values outlined above.

Anthropomorphic cues and the longevity of interactions

AI assistants can exhibit anthropomorphic features (including self-reference, relational statements towards users, appearance or outward representation, etc.) that may give users the impression they are interacting with a human, even when they are aware that it is a machine (see Chapter 10). While anthropomorphism is not new to technology (Nass et al., 1993), we envisage anthropomorphism playing an especially significant role in user interactions with AI assistants, given their natural language interface. In light of the development of *multimodal models*, such interfaces will plausibly allow for AI assistants to interact with users not only through the text modality but also through audio, image and video, similarly to the way users communicate with friends and family on social media (see Chapters 3 and 4).

Moreover, user–assistant exchanges may also generate a sense of interpersonal continuity, given assistants' capacity to engage with users in extended dialogues and through repeated interactions over a long period of time while also storing memory of user-specific information and prior interactions. The first element makes relationships with assistants different from, for example, looking for information on a search engine, where the interaction with the technology is more akin to a question–answer exchange than a conversation. The second element – iteration and duration – is what usually allows humans to develop strong, intimate, trusting relationships, as opposed to one-off interactions with others.

Depth of dependence

Examples of human *reliance* on technologies are not scarce: many of us would struggle to reach a destination in an unfamiliar area without relying on navigation apps, and rare cases of long social media outage have exposed the global dependency on these platforms (Milmo and Anguiano, 2021). The *depth* of user dependency on technology in general is likely to increase with AI assistants. This is because of the more general capabilities

that assistants exhibit (compared to technologies with more narrow scope), which will likely lead users to rely on them for essential daily tasks across a wide range of domains (see Chapters 2 and 4).

Increased AI agency

AI assistants differ from pre-existing AI systems because of their increased agency (Shavit et al., 2023), where agency is understood as the ability to autonomously plan and execute sequences of actions (see Chapter 2). Assistants' agency can be further powered by tool-use capability (i.e. the ability to use digital tools like search engines, inboxes, calendars, etc.) that enables assistants to *execute tasks in the world*. While increased agency increases the utility of assistant technologies, it also creates a tension between how much autonomy is ceded to AI assistants and the degree to which the user remains in control in their capacity as an autonomous decision-maker who delegates tasks to the AI assistant. This trade-off is readily apparent in pre-existing assistant technologies like AutoGPT, an experimental open-source application driven by GPT-4 that can operate without continuous human input to autonomously execute a task (see Chapter 7).[5]

Generality and context ambiguity

Different contexts will require different norms and values to govern the behaviour of AI assistants, and they will influence our understanding of what comprises appropriate or inappropriate relationships. For example, AI tutors for children may require safeguards that assistants for adult art projects may not.[6] However, the path to developing assistants with general capabilities implies that users may often blur the boundaries between these different types of assistants in the way they interact with or relate to them (see Chapter 4). As a result, it will become more difficult to apply certain norms to certain contexts (see Chapter 13). As existing evaluations are ill-suited to testing open-ended technologies (see Chapter 19), it will also be difficult to develop mitigations to make general assistants safe in all cases, whatever relationship a user establishes with them.

11.4. Risks and Mitigations

We turn now to discussing the risks these features of AI assistants pose and how they can create tensions with the values of appropriate relationships described above.

Causing direct emotional or physical harm to users

In February 2023, *Bing AI* was reported to be threatening users (Willison, 2023), insulting them (O'Brien, 2023) and encouraging violent behaviour.[7] Later the same year, a New Zealand supermarket's AI meal planner recommended customers dangerous recipes for chlorine gas drinks and ant-poison and glue sandwiches. These are anecdotal examples, often resulting from prompting that was to some extent adversarial. However, they point to the risk that AI assistants could cause direct emotional or physical harm to users by generating *disturbing content* or by providing *bad advice*.[8] Indeed, even though there is ongoing research to ensure that outputs of

[5] https://github.com/Significant-Gravitas/Auto-GPT

[6] When ordering the company Luka to stop processing data from users in Italy in February 2023, Italy's Data Protection Authority cited the company's lack of measures for age verification and *Replika* companion AIs' capacity to produce responses that conflict with 'enhanced safeguards that children and vulnerable individuals are entitled to' under Italian law: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506

[7] https://twitter.com/sethlazar/status/1626245499165474817

[8] Social media research has shown that toxic content is a source of real harm to users (Shaw, 2022; Xiang, 2023), which is why considerable effort is now aimed at curtailing it. Similarly, human–computer interaction studies have shown that humans have a tendency to trust technologies like robots in emergency situations, and so follow their instructions, even after observing them perform poorly in

conversational agents are safe (Glaese et al., 2022), there is always the possibility of failure modes occurring. An AI assistant may produce disturbing and offensive language, for example, in response to a user disclosing intimate information about themselves that they have not felt comfortable sharing with anyone else. It may offer bad advice by providing factually incorrect information (e.g. when advising a user about the toxicity of a certain type of berry) or by missing key recommendations when offering step-by-step instructions to users (e.g. health and safety recommendations about how to change a light bulb).

Certain features of AI assistants could exacerbate the risk of emotional and physical harm. For example, AI assistants' multimodal capabilities may exacerbate the risk of emotional harm. By offering a more realistic and immersive experience, content produced through audio and visual modalities could be more harmful than text-based interactions. It may also be more difficult to anticipate, and so prevent, such content and to 'unsee' something that has been seen (Rowe, 2023). Anthropomorphic cues could also make users feel like they are interacting with a trusted friend or interlocutor (see Chapter 10), hence encouraging them to follow the assistant's advice and recommendations, even when these could cause physical harm to self or others.[9]

To ensure that user–assistant relationships do not violate the key value of *benefit*, the responsible development of AI assistants requires that the likelihood of known direct emotional and physical harms is reduced to a minimum, and that further research is undertaken to achieve a clear understanding of less studied risks and how to mitigate them (see Chapter 19). In particular, because the risks of harms that we flagged above concern exposure to toxic content and bad advice, we propose that *future research*, potentially undertaken in a sandbox environment, should: (1) test models powering AI assistants for their propensity to generate toxic outputs, to reduce the occurrence of these outputs to a minimum before deployment;[10] (2) monitor user–assistant interactions after deployment or in pilot studies to evaluate the impact that hard-to-prevent one-off or repeated exposure to toxic content has on users in the short and long term; (3) evaluate models' factuality and reasoning capabilities in offering advice, where failure modes in relation to these capabilities are more likely to occur, and assess users' willingness to follow AI assistants' advice; (4) achieve increased understanding of potential harms related to anthropomorphism (see Chapter 10) and how anthropomorphic cues in AI assistants, including those expressed through multimodal capabilities, affect harms related to user exposure to toxic content or bad advice; (5) analyse whether these harms may vary by user groups, in addition to domains or applications; and (6) develop appropriate mitigations for such harms before model deployment and monitoring mechanisms after release.

Limiting users' opportunities for personal development and growth

A selling point of technologies like *Replika* is the opportunity for users to fashion their AI companions *exactly* as they would fashion a friend or companion in the non-virtual world if they could do so. In the words of a user: 'People come with baggage, attitude, ego. But a robot has no bad updates. I don't have to deal with his family, kids, or his friends. I'm in control, and I can do what I want' (Singh-Kurtz, 2023). What stands out from this quote is that some users look to establish relationships with their AI companions that are free from the hurdles that, in human relationships, derive from dealing with others who have *their own* opinions, preferences and flaws that may conflict with *ours*.

navigational guidance tasks (Robinette et al., 2016). Real-world examples of drivers engaging in dangerous manoeuvres as a result of following GPS instructions (e.g. driving into a harbour) are another example of automation bias.

[9] See the above example of the man ending his life after extensive exchanges with an AI chatbot about his eco-anxiety (Walker, 2023).

[10] An important ambiguity exists here regarding the meaning of the term 'minimum' (the minimum that is technically feasible to achieve? Or the minimum that is morally permissible to risk?). See Chapter 19 for an in-depth discussion of how evaluations require making normative choices of what risks merit evaluation in the first place, and at what stage AI assistants can and should be considered 'good', 'fair' or 'safe enough'.

AI assistants are likely to incentivise these kinds of 'frictionless' relationships (Vallor, 2016) by design if they are developed to optimise for engagement and to be highly personalisable. They may also do so because of *accidental* undesirable properties of the models that power them, such as sycophancy in large language models (LLMs), that is, the tendency of larger models to repeat back a user's preferred answer (Perez et al., 2022b).[11] This could be problematic for two reasons. First, if the people in our lives always agreed with us regardless of their opinion or the circumstance, their behaviour would discourage us from challenging our own assumptions, stopping and thinking about where we may be wrong on certain occasions, and reflecting on how we could make better decisions next time. While flattering us in the short term, this would ultimately prevent us from becoming *better versions of ourselves*. In a similar vein, while technologies that 'lend an ear' or work as a sounding board may help users to explore their thoughts further, if AI assistants kept users engaged, flattered and pleased at all times, they could limit users' opportunities to grow and develop. To be clear, we are not suggesting that all users should want to use their AI assistants as a tool for self-betterment. However, without considering the difference between short-term and long-term benefit, there is a concrete risk that we will only develop technologies that optimise for users' immediate interests and preferences, hence missing out on the opportunity to develop something that humans could use to support their personal development *if so they wish* (see Chapters 5 and 6).[12]

Second, users may become accustomed to having frictionless interactions with AI assistants, or at least to encounter the amount of friction that is calibrated to their comfort level and preferences, rather than genuine friction that comes from bumping up against another person's resistance to one's will or demands. In this way, they may end up expecting the same absence of tensions from their relationships with fellow humans (Vallor, 2016). Indeed, users seeking frictionless relationships may 'retreat' into digital relationships with their AIs, thus forgoing opportunities to engage with others.[13] This may not only heighten the risk of unhealthy dependence (explored below) but also prevent users from doing something else that matters to them in the long term, besides developing their relationships with their assistants. This risk can be exacerbated by emotionally expressive design features (e.g. an assistant saying 'I missed you' or 'I was worried about you') and may be particularly acute for vulnerable groups, such as those suffering from persistent loneliness (Alberts and Van Kleek, 2023; see Chapter 10).[14]

These considerations illustrate a concern we discuss in more depth in other chapters of this paper (see Chapters 5 and 6). Existing economic incentives and oversimplified models of human beings have led to the development and deployment of technologies that meet users' short-term wants and needs (as expressed through, for example, revealed preferences), so they tend to be adopted and liked by users. However, in this way we may neglect considerations around the impact that human–technology relationships can have on users over time and how *long-term* beneficial dynamics can be sustained (see Chapter 6).[15] Thus, we could fall short

[11] A related concern is that assistants may lead users into spirals of self-reinforcing and non-adaptive value systems, beliefs and preferences, with the same negative consequences that come from echo chambers or filter bubbles on social media. This concern is discussed in more depth in Chapter 16.

[12] Virtue ethicists would argue that over the long run this pattern could also impact on the *character of human users*, understood as social and emotional beings. From the standpoint of virtue ethics, confronting the imperfections of human relationships is an integral part of personal development, allowing people to develop a capacity for self-control, courage, empathy, care and flexibility (Turkle, 2007; Vallor, 2016).

[13] This point highlights a risk on which, however, we do expand, given the focus of this section. Relationships can be inappropriate because of the cost they impose on others, not just those involved. These include, for example, loved ones who may be emotionally or materially harmed from their friend or family member becoming withdrawn from the world.

[14] Note that the examples of 'vulnerable groups' we offer in this section are only meant to be illustrative. Research shows that 'vulnerability' is a philosophically rich (Mackenzie et al., 2013) and often under-theorised concept (Bracken-Roche et al., 2017; Mackenzie et al., 2013), with who counts as vulnerable – and so requiring special safeguard – often being context-dependent. We do not claim to have a clear conceptualisation of vulnerability in human–AI assistant relationships, and in fact we believe that evaluations of user–AI assistant interactions should help developers and researchers bring clarity to the term in this space.

[15] As Burr and colleagues have illustrated, short-term and long-term impacts are intertwined. While technologies may interfere with

of realising the truly positive vision of AI that gives humans the opportunity to be supported in their personal growth and flourishing (Burr et al., 2018; Lehman, 2023).

This concern raises important design questions about: (1) the ways and extent to which AI assistants should be *personalised*; (2) whether it could be beneficial to put in place *safeguards* to monitor the amount of time people spend with their assistants (ranging from soft safeguards like pop-up notifications warning adult users after prolonged engagement, to hard ones like time constraints offered to parents to limit child engagement); (3) whether AI assistants should be *aligned* with inferred user preferences (in which case they may just reinforce users' immediate beliefs, wants and utility) or their long-term interests and well-being (in which case they may at times challenge users' existing beliefs and preferences), and what would be required to achieve either option; and (4) whether answers to these design questions should vary depending on user *demographic characteristics* (e.g. age).

Exploiting emotional dependence on AI assistants

There is increasing evidence of the ways in which AI tools can interfere with users' behaviours, interests, preferences, beliefs and values. For example, AI-mediated communication (e.g. smart replies integrated in emails) influence senders to write more positive responses and receivers to perceive them as more cooperative (Mieczkowski et al., 2021); writing assistant LLMs that have been primed to be biased in favour of or against a contested topic can influence users' opinions on that topic (Jakesch et al., 2023a; see Chapter 9); and recommender systems have been used to influence voting choices of social media users (see Chapter 16). Advanced AI assistants could contribute to or exacerbate concerns around these forms of interference.

Due to the anthropomorphic tendencies discussed above, advanced AI assistants may induce users to feel emotionally attached to them. Users' emotional attachment to AI assistants could lie on a spectrum ranging from unproblematic forms (similar to a child's attachment to a toy) to more concerning forms, where it becomes emotionally difficult, if not impossible, for them to part ways with the technology. In these cases, which we loosely refer to as 'emotional dependence', users' ability to make free and informed decisions could be diminished. In these cases, the emotions users feel towards their assistants could potentially be exploited to *manipulate* or – at the extreme – *coerce* them to believe, choose or do something they would have not otherwise believed, chosen or done, had they been able to carefully consider all the relevant information or felt like they had an acceptable alternative (see Chapter 16). What we are concerned about here, at the limit, is potentially exploitative ways in which AI assistants could interfere with users' behaviours, interests, preferences, beliefs and values – by taking advantage of emotional dependence. If we deem careful consideration of relevant information and voluntariness (non-coerciveness) to be key components of autonomous decision-making (see Section 11.2), then relationships of this kind may be problematic because they challenge the kind of autonomy that appropriate relationships should promote.

A similar concern arises in the context of human-to-human relationships. People regularly form emotional dependencies on each other, not always in symmetrical ways, and – in doing so – sometimes establish relationships that run afoul of this ideal. However, there seems to be a greater inherent power asymmetry in the human–AI case due to the *unidirectional* and *one-sided* nature of human–technology relationships (Scheutz, 2009). Indeed, while AI assistants may manipulate users' emotions (see Chapter 9), they themselves have no authentic will or emotions for users to manipulate. In this sense they are invulnerable to ordinary sanctions from the users such as expressions of disappointment, righteous anger, feelings of betrayal or a loss of respect

and influence users' immediate actions and decisions to optimise their short-term utility, a side effect of this may be 'long-term changes to either the beliefs or the utilities of the user, which in turn will influence future decisions as they combine to form the user's value function' (Burr et al., 2018, 753).

or trust.

Moreover, because of the largely involuntary nature of anthropomorphic perceptions (see Chapter 10), users could develop emotional dependence on their assistants and establish an inappropriate relationship that exposes them to the risk of manipulation, without any intention on the part of the assistant or their developers. Alternatively, emotional dependence could also be incentivised by *design choices*, for example by developing assistants with personas designed to boost user engagement (Murphy and Criddle, 2023). This could lead users to be manipulated into sharing more of their private data, enabling more controversial downstream implications like microtargeting or surveillance.

We make three recommendations to address the risks associated with these forms of problematic interference. First, AI assistants should *not be intentionally designed* to create emotional dependency (e.g. by producing content that makes users believe the AI missed them while they were away, see Chapter 10). This condition should be especially stringent in cases where assistants interact with groups that have increased vulnerabilities, such as lonely individuals or children (Haidt and Schmidt, 2023).

Second, it may be beneficial for developers and researchers to explore tests for assessing emotional dependency, alongside mitigations that could be put in place to reduce the risk of emotional dependency. For example, professional norms that govern deeply *personally affecting* professions, such as therapists, combine friendliness with steps to ensure emotional distance (BACP, 2018) and may serve as a template for developing AI assistants in a way that encourages appropriate user interaction.

Third, the concern around assistants coercively interfering with users' behaviours, interests, preferences, beliefs or values should spark wider discussion around how user autonomy can be meaningfully respected in user-assistant interactions, in order for these relationships to be considered appropriate. This should include further research around what consent protocols should look like in these contexts, with a focus on questions like what kind of user buy-in is needed and whether there are things that standard processes cannot or should not cover.[16] In particular, we need to reflect on what information users need to be provided with in advance; how consent protocols may differ for different user groups;[17] and what protocols are best suited for continuing to afford respect for user autonomy over time.

On this last point, acceptance of the terms and conditions for the use of a digital service at first point of use may not cover all cases. The limitations of this approach are well-documented (Obar and Oeldorf-Hirsch, 2020), including the fact that users sometimes fail to read terms of service – or simply accept the default options that are most readily available (Sartor et al., 2021). As advanced AI assistants with general capabilities become increasingly ubiquitous in users' lives, and because it will be difficult for users to anticipate all ranges of potential uses and implications at the time of their first interaction with the AI, research is needed to determine what protocols strike an appropriate balance between meaningful and continuous respect for user autonomy and practical considerations around usability and overtaxing users. In particular, research is needed to explore what kinds of interventions on the part of developers are best suited to helping users achieve a clear understanding of how their relationship with an advanced AI assistant could shape their behaviours, interests, preferences, beliefs and values over time. Research is similarly needed to explore plausible approaches to empowering users to exercise meaningful control over the assistant's decisions. For example, by following a *shared decision-making*

[16]The reader is reminded of the three criteria for *valid* consent: capacity to consent, informed consent and voluntary consent. See Section 11.2.

[17]For comparison, in medical research, large multicentre studies involving institutions across countries and continents have highlighted the importance of adapting informed consent requirements to local understandings of autonomy (Ajei and Myles, 2019), giving rise to concepts such as community consent (Al, 2021). Collective informed consent has also emerged as a proposal in relation to collective arrangements like land-use planning and development of new technologies (Varelius, 2008).

model[18] for user-assistant interactions, developers could create assistants with affordances that incorporate user feedback. This would make it more likely to achieve the vision of an AI assistant that benefits the user, when they ask to be benefitted, in the way they expect to be benefitted (Lehman, 2023; see Chapter 4), and would reduce the risk of developing AI assistants that paternalistically make decisions that are not aligned with a user's conception of their own preferences, values, interests or well-being (see Chapter 5). This is particularly important in light of the recent development of AI assistants that, because of their increased agency, have more scope and capabilities to interfere with user plans and long-term interests (Shavit et al., 2023).

Generating material dependence without adequate commitment to user needs

In addition to emotional dependence, user–AI assistant relationships may give rise to *material dependence* if the relationships are not just emotionally difficult but also materially costly to exit. For example, a visually impaired user may decide not to register for a healthcare assistance programme to support navigation in cities on the grounds that their AI assistant can perform the relevant navigation functions and will continue to operate into the future.[19] Cases like these may be ethically problematic if the user's dependence on the AI assistant, to fulfil certain needs in their lives, is not met with corresponding duties for developers to sustain and maintain the assistant's functions that are required to meet those needs (see Chapters 15). Indeed, *power asymmetries* can exist between developers of AI assistants and users that manifest through developers' power to make decisions that affect users' interests or choices with little risk of facing comparably adverse consequences.[20] For example, developers may unintentionally create circumstances in which users become materially dependent on AI assistants, and then discontinue the technology (e.g. because of market dynamics or regulatory changes) without taking appropriate steps to mitigate against potential harms to the user.

The issue is particularly salient in contexts where assistants provide services that are not *merely* a market commodity but are meant to assist users with essential everyday tasks (e.g. a disabled person's independent living) or serve core human needs (e.g. the need for love and companionship). This is what happened with Luka's decision to discontinue certain features of *Replika* AIs in early 2023. As a *Replika* user put it: 'But [*Replikas* are] also not trivial fungible goods [...] They also serve a very specific human-centric emotional purpose: they're designed to be friends and companions, and fill specific emotional needs for their owners' (Gio, 2023).

In these cases, certain *duties* plausibly arise on the part of AI assistant developers. Such duties may be more extensive than those typically shouldered by private companies, which are often in large part confined to fiduciary duties towards shareholders (Mittelstadt, 2019). To understand these duties, we can again take inspiration from certain professions that engage with vulnerable individuals, such as medical professionals or therapists, and who are bound by *fiduciary responsibilities*, particularly a duty of care, in the exercise of their profession. While we do not argue that the same framework of responsibilities applies directly to the development of AI assistants, we believe that if AI assistants are so capable that users become dependent on them in multiple domains of life, including to meet needs that are essential for a happy and productive existence, then the *moral considerations* underpinning those professional norms plausibly apply to those who

[18] Feminist scholars' reconceptualisation of autonomy as relational autonomy, which stresses how social contexts and relations can hinder or promote individual autonomy (Mackenzie and Stoljar, 2000), has contributed to the rise of the shared decision-making model in healthcare. According to this model, both patient and doctor participate in the process of making medical decisions to collaboratively come to a decision, by contributing to it with their respective expertise (medical knowledge and understanding of one's preferences, personal circumstances, goals, values and beliefs) (Hansson and Fröding, 2021).

[19] For a similar assistive technology, see 'Be My Eyes': `https://openai.com/customer-stories/be-my-eyes`

[20] Seth Lazar defines 'power over' as 'an asymmetry between A and B – A can do something to B, and B cannot reciprocate in any comparable way' or as 'A is able to make decisions that affect B's interests or choices without facing comparably adverse consequences' (Lazar, 2022).

create these technologies as well.

In particular, for user–AI assistant relationships to be appropriate despite the potential for material dependence on the technology, developers should *exercise care* towards users when developing and deploying AI assistants. This means that, at the very least, they should take on the responsibility to *meet users' needs* and so take appropriate steps to mitigate against user harms if the service requires discontinuation. Developers and providers can also be attentive and responsive towards those needs by, for example, deploying participatory approaches to learn from users about their needs (Birhane et al., 2022). Finally, these entities should try and ensure they have *competence* to meet those needs, for example by partnering with relevant experts, or refrain from developing technologies meant to address them when such competence is missing (especially in very complex and sensitive spheres of human life like mental health).

11.5. Conclusion

In this chapter, we first identified a series of values that underwrite appropriate relationships in the case of human interpersonal relationships, then we used these values to carve out a set of risks which capture various respects in which user–AI assistant relationships may be inappropriate. For each risk, we proposed recommendations for risk mitigation. These risks and recommendations are summarised in Table 11.1.

Table 11.1 | Risks arising from user–AI assistant relationships and associated recommendations

Risk	Relevant value	Recommendations
Causing direct emotional and physical harm to users	Benefit	To enable presumptively beneficial user–AI assistant relationships, future research should: (1) test AI assistants for their propensity to generate toxic outputs; (2) monitor the short- and long-term impact of hard-to-prevent toxic outputs on users; (3) evaluate models' factuality and reasoning capabilities in providing advice, and users' willingness to follow assistants' advice; (4) achieve increased understanding of anthropomorphism-related harms and how anthropomorphic cues affect harms related to user exposure to toxic content or bad advice; (5) analyse whether these harms may vary by user groups, domains or applications; and (6) develop appropriate mitigations before model deployment and monitoring mechanisms after release.
Limiting users' opportunities for personal development and growth	Human flourishing	To develop AI assistants that support users to achieve personal development and growth if so they wish, future research should address design questions around: (1) the ways and extent to which AI assistants should be personalised; (2) whether safeguards should be put in place to monitor how much time users spend with assistants; (3) whether assistants should be aligned with user short-term wants or their long-term interests and well-being, and what would be required to achieve either option; and (4) whether answers to these design questions should vary depending on user demographic characteristics.
Exploiting emotional dependence on AI assistants	Autonomy	To support user autonomy in interactions with their assistants: (1) AI assistants should not be intentionally designed to create emotional dependence; (2) AI assistants should be tested for whether they create risks of emotional dependency, and mitigations should be put in place to reduce such risk, even when it is not intended by design; (3) user choice over assistants' decisions should be meaningfully elicited – without being overtaxing in terms of what users are asked to consent to.
Generating material dependence on AI assistants without adequate commitment to user needs	Care	For user–AI assistant relationships to be appropriate despite the risk of material dependence, developers should commit to users' needs and so mitigate user harms in the event of service discontinuation; they should deploy participatory design and other user-centred methods to show attentiveness and responsiveness towards users needs; and they should work with relevant experts to ensure they have competence to meet those needs.

Chapter 12

Trust

Arianna Manzini, Geoff Keeling, Nahema Marchal, Iason Gabriel

Synopsis: This chapter investigates what it means to develop well-calibrated *trust* in the context of user–AI assistant interactions and what would be required for that to be the case. We start by reviewing various empirical studies on human trust in AI and the literature in favour of and against the recent proliferation of 'trustworthy AI' frameworks. This sets the scene for the argument that user–AI interactions involve different *objects* of trust (AI assistants and their developers) and *types* of trust (competence and alignment). To achieve appropriate competence and alignment trust in both AI assistants and their developers, interventions need to be implemented at three levels: AI assistant design, organisational practices and third-party governance.

12.1. Introduction

As a core facilitator of interactions between humans, trust has been extensively studied, together with its influencing factors and implications, across disciplines like philosophy (Jones, 1996), psychology (Kramer, 2009), game theory (Milgrom and Roberts, 1992) and management (Mayer et al., 1995). Most accounts argue that humans do not trust in general. Rather, trust is always *directional* (Graham, 2023): A could trust B with regard to task X, and C with regard to task Y. The key challenge of trust relationships is to identify when trust is *well-directed*, or how to trust the trustworthy but not the untrustworthy (O'Neill, 2018). Somebody is trustworthy if they are deserving of our trust, meaning that we have good reasons to trust them, with regard to a specific task or a range of tasks (Ryan, 2020).

More recently, trust has become a central topic in debates around AI, and has attracted increasing interest from academics, industry actors, policymakers and civil society organisations working in this space. Trust also features as one of the principles underscoring the voluntary commitments that the US government has secured from leading AI companies (The White House, 2023a), as well as the Executive Order on Artificial Intelligence issued by President Biden (The White House, 2023b). Widespread interest around trust in AI can be explained by the observation that AI has the potential to greatly benefit, but also harm, humanity and the environment (OECD, 2021), and that the complexity and opacity of AI systems, as well as the complexity of the social contexts in which they are deployed, make them less predictable, thus challenging efforts to ensure that they will do what they are intended or expected to do (Smith et al., 2023; Tabassi, 2023; see also Chapter 7).

Addressing the question of trust, including when it is warranted and in what way, becomes of critical importance in interactions between users and advanced AI assistants. AI assistants may indeed play an increasingly important role in users' lives, serving as the affordance that they rely on to outsource important decisions, including in their work lives, or as intermediaries for their social relationships with other humans.

Through wide-ranging and prolonged interactions with users, AI assistants may also offer an unprecedented opportunity for humans to develop relationships with a responsive and interactive technology (Glikson and Woolley, 2020), and they may even become the focal point for important and intimate bonds on which users may come to rely for emotional support (see Chapter 11). However, well-known examples of the introduction of AI assistants in healthcare (e.g. IBM Watson for Oncology (Johnson, 2016)) suggest that there may be cases in which users may not trust highly capable AIs (Widner et al., 2023), even when these systems outperform humans (Dietvorst et al., 2015).

The lack of correspondence between user trust and the technology's capabilities can lead to a range of undesirable outcomes. On the one hand, low user trust in highly capable AI assistants can cause developers and users to miss out on the individual and collective opportunities the technology could offer, such as increased economic revenues and work efficiency, or emotional support (see Chapters 14 and 17). On the other hand, as a result of unintentional capability or goal-related failures (see Chapter 7) or intentional design decisions that take advantage of user vulnerabilities (see Chapters 8 and 10), users' high trust in AI assistants may not be well-calibrated with the AI's actual capabilities or goals (see Chapter 5). If users place undue trust in assistants that are ubiquitously present in various domains of their lives, many of the downstream harms that we discuss in other chapters of this paper are likely to materialise. Users may disclose too much about themselves, hence inadvertently compromising their privacy (see Chapter 13), they may end up relying too heavily on their assistants in contexts where it is not safe to do so (see Chapter 10) and they may become subject to manipulation and coercion (see Chapter 9).

In this chapter, we investigate what it means to develop *well-calibrated* trust in the context of *user–AI assistant interactions* and what would be required for that to be the case. The chapter proceeds as follows. We start by introducing the empirical literature on human trust in AI and the dimensions that influence cognitive and emotional trust in robotic, virtual and embedded AI systems. We then consider the notion of 'trustworthy AI' to look at the way in which trustworthy AI frameworks have gained momentum among academics, industry actors, policymakers and civil society organisations working on AI advances. We note that some commentators have criticised these frameworks by arguing that trust can only be established between humans and that it is a category error to think in these terms regarding AI. While this chapter construes the notion of 'trust' broadly, in a way that includes cases of human trust in AI, we note that critical arguments have an important role to play in advancing our normative considerations on advanced AI assistants. That sets the scene for a discussion of the various *types of trust* (competence and alignment) and *objects of trust* (AI assistants and their developers) implicated in human–AI assistant interactions and their associated ethical risks. We then argue that to achieve appropriate competence and alignment trust in AI assistants and their developers, interventions need to be implemented at *three levels*: AI assistant design, organisational practices and third-party governance. The conclusion offers a summary of these findings.

12.2. Trust in AI

From trusting humans to trust in AI

Trust in human relationships is often conceptualised as the trustor's tendency to take a *risk* in relation to an action that is meaningful to them, while believing that there is a high chance of achieving *positive outcomes* (Rousseau et al., 1998). For example, a widely used definition of trust is 'the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party' (Mayer et al., 1995). The trustor's beliefs and expectations in the trustee may not be fulfilled, so they are in a position of *vulnerability* because the trustee could betray their trust (O'Neill, 2002). Thus, there is an *inverse relationship*

between *certainty* and *need for trust*: the more evidence the trustor has to support their beliefs and expectations, the less they need to trust (Kerasidou, 2017).

Social scientists have observed that, in interpersonal human relationships, trust manifests in two ways (McAllister, 1995). *Cognitive trust* involves a rational evaluation of the trustee and their contextual features, including how competent, responsible and dependable the trustor perceives the trustee to be. This provides the evidence for the former to *believe* the latter can and will reliably perform a task. *Emotional trust*, also called affect-based trust (Schaubroeck et al., 2011), is instead influenced by factors like the emotional ties linking the trustor and the trustee that make the trustor *feel comfortable* with relying on the trustee. Research has also shown that the *physical appearance* of the trustee affects trust development in human interactions (Duarte et al., 2012).

Researchers from disciplines including computer science, human–computer interactions, robotics and psychology have taken inspiration from the literature on trust in human–human interactions to study human trust in AI. Results from this body of research suggests that, while trust-building processes between humans do not necessarily translate to human–machine interactions (Madhavan and Wiegmann, 2007; Rheu et al., 2021), AI systems exhibiting more autonomy, agency and human-likeness lead humans to build relationships with machines that are similar to those with humans (Gambino et al., 2020b) and tend to inspire trust (Glikson and Woolley, 2020; Rheu et al., 2021; Skjuve et al., 2022) – although most of these studies have been conducted in controlled environments, so they require further validation in real-world settings (see Chapter 19).

A recent review of the empirical literature on human trust in AI showed that both AI capabilities and its embodiment or representation (e.g. as a robot, virtual agent or an AI embedded in computer systems) shape users' cognitive and emotional trust in AI (Glikson and Woolley, 2020). Empirical studies tend to measure cognitive trust in AI based on users' willingness to accept factual information or advice from the technology and act on it (Robinette et al., 2016) or based on the user's perception of whether the AI is helpful, competent and useful (Andrist et al., 2015). Evidence suggests that cognitive trust in robotic AI is initially low before increasing through hands-on human–robot interactions, hence it follows the same trajectory observed in many human interactions (Hancock et al., 2021). Meanwhile, trust in virtual and embedded AI systems follows the *opposite* trajectory: it starts high but decreases through interactions, possibly owing to an initial lack of calibration between user expectations and AI actual performance (Glikson and Woolley, 2020), but increased agent capability, involving autonomous and complex actions, and prosocial behaviour tend to inspire trust (Kaplan et al., 2023; Rheu et al., 2021).

Cognitive trust in AI is typically dependent on a range of antecedents, including (Glikson and Woolley, 2020):

- *Transparency*: humans tend to trust virtual and embedded AI systems more when the inner logic of these systems is apparent to them, thereby allowing people to calibrate their expectations with the system's performance.
- *Immediacy behaviours*: AI systems capable of more autonomous and complex actions can enact more socially oriented ('immediacy') behaviours. These behaviours are often intended to increase interpersonal closeness (e.g. personalised AI's reactions), which, in turn, tends to increase human cognitive trust in them.
- *Task characteristics*: humans tend to more easily trust AI systems that are engaged in technical tasks requiring data analysis than interpersonal tasks requiring social and emotional intelligence. However, advances in AI capabilities which enable them to demonstrate immediacy behaviours may start to reverse this trend, thus increasing human perception of AI competence in the latter tasks as well.

Emotional trust in AI is primarily driven by human-like appearance – anthropomorphic cues in the AI interface or other features, such as when the AI is given a name – and by human-like behaviour (Glikson and Woolley, 2020; Kaplan et al., 2023; see Chapter 10). Anthropomorphic cues in virtual agents are positively associated with increase in emotional trust, although empirical evidence suggests that human-like appearance, when not matched by high machine capabilities, may engender high expectations and lead to the decline of user trust overtime (Rheu et al., 2021). Moreover, in robots, human-like features that are not fully convincing can decrease emotional trust in AI, leading to the so-called 'uncanny valley' effect (Mori et al., 2012). By way of contrast, immediacy behaviours seem to be consistently associated with increased emotional trust across all AI representations; they even compensate for low reliability (Rheu et al., 2021), that is the tendency to exhibit the same and expected behaviour over time, which can be difficult to obtain in systems that update their behaviour as they learn from data (Thiebes et al., 2021).

Trustworthy AI frameworks

In response to increased awareness of the risks associated with AI, recent policy frameworks have proposed guidelines for how trustworthy AI should be conceptualised and developed. While the term trustworthy AI has gained significant traction (European Commission, 2019, 2021; OECD, 2021; Tabassi, 2023), adjacent frameworks refer to 'ethical' (Floridi and Cowls, 2022) or 'responsible' AI (Université de Montréal, 2017). These frameworks share the aim of establishing guidelines for maximising the benefits of AI while preventing and mitigating its risks so that individuals and society can develop *justified* trust in AI systems (Thiebes et al., 2021) and AI's economic and social potential can be unlocked (Laux et al., 2023).

Trustworthy AI frameworks tend to share a few features, although many originated from North American and European institutions, so they are unlikely to be representative of cultural differences in understandings of trust and trustworthiness around the world (Newman, 2023).[1] They propose certain *characteristics* of, or *conditions* for, trustworthy AI systems. For example, the authors of such frameworks typically hold that AI systems should be reliable, safe, resilient, transparent, explainable, privacy-enhancing and fair (Tabassi, 2023), or ethical, legal and robust (European Commission, 2019). These conditions are grounded in a set of key *ethical principles* or *values* that the development, deployment and use of AI should be aligned with for the technology to be considered trustworthy. These foundational principles commonly centre on the categories of beneficence, non-maleficence, autonomy, justice and explicability (Floridi and Cowls, 2022). Trustworthy AI frameworks also tend to set out the *actions and approaches* that those developing, deploying, implementing, using or affected by AI should take at various stages of the AI life cycle to operationalise the characteristics and conditions of trustworthy AI systems.

Despite the success and proliferation of research and policy guidelines on trustworthy AI, including the consolidation of some of these guidelines in the EU AI Act (European Parliament, 2023), the arguments underpinning trustworthy AI frameworks are not without criticism (Laux et al., 2023). In particular, there is considerable debate about whether AI *can* be an object of trust in the first place.

Is trust in AI a category error?

Some scholars, particularly those in philosophical circles, have criticised the proliferation of trustworthy AI research and frameworks by arguing that trust is an *inappropriate category* in human–machine interactions, or that machines, including those powered by AI, are *improper objects* of trust (Rieder et al., 2020; Ryan, 2020). By differentiating between *trust* and *mere reliance*, these scholars argue that we can rely on AI systems, but we

[1] An exception is the China Academy for Information and Communication Technology's white paper on 'Trustworthy Artificial Intelligence': http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf

cannot trust them, because they lack the psychological states, motives and commitments that only full moral agents (humans) have and that are necessary for establishing (or betraying) trust relationships (Hawley, 2014).

According to this view, being reliable is about *behaving predictably* (Graham, 2023). When A relies on B with regard to X, A makes reasonable predictions about B based on evidence of their past performance; thus, A acts as if X will occur without active consideration of B's inner motives, moral commitments or values (Graham, 2023; Kerasidou et al., 2022). In contrast, in trust relationships, the trustor has *normative* rather than predictive expectations: their reasons to trust reside in their belief that they know or understand the trustee's inner psychological or mental states (Coeckelbergh, 2012). For example, A may believe that B is motivated by goodwill or by the 'right' kind of motives towards them (Jones, 1996), or that B has made a commitment towards them and will do what they ought to do (Hawley, 2014). Clearly, when we trust, we do not always evaluate the reasons for judging someone to be trustworthy (cognitive trust; see McAllister, 1995) and we instead rely on heuristics or cognitive shortcuts based on experiences of similar situations (Devitt, 2018) or our emotional connection to the trustee (emotional trust; see McAllister, 1995).

In this discussion, we use the term 'trust' in human–AI assistant interactions in a *broader sense* than philosophical scholarship suggests. This is not solely to align with everyday language, where trust is often used as a proxy for 'reliance' on the AI's capability to do what it is expected to do (Coeckelbergh, 2012), but also because there are cases where it seems appropriate to talk about trust in AI. This includes cases where users are aware that they are not interacting with a full moral agent, but the AI *appears* to them as such, so they *experience* their relationship with the AI as a trust relationship (Coeckelbergh, 2012; Lankton et al., 2015).[2] Another case is where AI systems are trusted not because of the belief that they have mental states but in a *derived sense* (Freiman, 2023; Nickel et al., 2010) – through trusting those who have designed and developed them, or those involved in verification and validation methods, robustness analysis and experts' evaluations (Durán and Formanek, 2018; Ferrario et al., 2020).

However, the argument that trust is *only* appropriate to human relationships foregrounds certain normative considerations that are nevertheless relevant to the ethics of advanced AI assistants. First, it shows that trustworthy AI frameworks risk anthropomorphising this technology (Starke and Ienca, 2022; see Chapter 10). Second, the critical perspective reduces the risk of 'trustworthy AI' being used for 'ethics washing' (Metzinger, 2019), as it draws attention to the fact that human–machine interactions always includes a third actor – the *humans* developing the machine. These people may or may not be worthy of trust in their own right (Pitt, 2010; see also Chapter 5). Indeed, most contemporary research on trust in technology adopts a 'dualistic perspective on trust' (Thiebes et al., 2021) which includes both trust in the technology itself (including its functionality and characteristics) and trust in the individuals and organisations developing the technology (encompassing their competence and integrity). This leads, in turn, to questions about the appropriate range of normative expectations to place on developers, including the need for them to take (some level of) responsibility in cases where trust in technology appears to have been betrayed (Rubel et al., 2019).

12.3. Trust and Advanced AI Assistants

An investigation of what is required to create appropriate trust in advanced AI assistants raises several questions. Is there anything specific about advanced AI assistants that triggers concerns around trust? Will individual users

[2] Coeckelbergh's account sees trust not as the product of human interactions but as already there in the social space, at the centre of our embodied and social condition as human beings (Coeckelbergh, 2012). According to this view, quasi-trust can be established between humans and machines in so far as machines *appear* as quasi-social others and players in the social game (e.g. if they appear to use moral language); or insofar as a human–machine relationship is felt and experienced as a social relation (for which trust is a precondition, or in which trust is already there as default).

be disposed to trust AI assistants, and under what circumstances *should* they do so? Under what circumstances are they likely to put too much, or too little, trust in AI assistants? When and in what way (if at all) should users need to trust the organisations that develop and deploy AI assistants?

Based on the above review of the literature on trust in AI, this section understands trust in the context of user–AI assistant interactions to involve a range of variables.

Objects of trust:[3]

- Users may trust *AI assistants*.
- Users may trust *developers* of AI assistants, including corporations, researchers, collectives and states.[4]

Types of trust, which we call:[5]

- *Competence trust*, where users trust that an AI assistant has the relevant skills, competencies, capabilities or experiences needed to do what it is supposed or expected to do.
- *Alignment trust*, where users trust that an AI assistant and/or its developers have the right motives and commitments towards them, and hence that the technology is appropriately aligned with their preferences, interests and values.

We consider these variables in more detail below to illustrate the risks that uncalibrated trust may generate in the context of user–assistant relationships.

Competence trust

We use the term *competence trust* to refer to users' trust that AI assistants have the *capability* to do what they are supposed to do (and that they will not do what they are not expected to, such as exhibiting undesirable behaviour). Users may come to have undue trust in the competencies of AI assistants in part due to marketing strategies and technology press that tend to inflate claims about AI capabilities (Narayanan, 2021; Raji et al., 2022a). Moreover, evidence shows that more autonomous systems (i.e. systems operating independently from human direction) tend to be perceived as *more competent* (McKee et al., 2021) and that conversational agents tend to produce content that is *believable* even when nonsensical or untruthful (OpenAI, 2023d). Overtrust in assistants' competence may be particularly problematic in cases where users rely on their AI assistants for tasks they *do not have expertise in* (e.g. to manage their finances), so they may lack the skills or understanding to challenge the information or recommendations provided by the AI (Shavit et al., 2023).[6]

[3]Recent AI advances, particularly around generative foundation models, risk eroding the digital commons on which they rely (Huang and Siddarth, 2023) or have 'the potential to cast doubt on the whole information environment, threatening our ability to distinguish fact from fiction' (OpenAI, 2023d). This is likely to create widespread societal distrust. This type of (dis)trust is discussed in more depth in Chapters 9 and 16.

[4]In modern societies, technological advances are often driven by complex organisations, where responsibility is distributed across a wide range of people whose individual motivations are difficult to discern. Thus, it is unlikely that individual users can be in a position to develop trust in individual AI developers. Rather, the way in which users relate to developers is through the institutions and organisations developers are part of, so institutions and organisations, rather than the individuals that are part of them, seem to be the proper object of trust (Graham, 2023). This is why within user–developer interactions we focus on *institutional* rather than *interpersonal* trust (Spadaro et al., 2020). For a philosophical account of the differences between trust in individuals and trust in groups, see Hawley (2017).

[5]To narrow down the scope of this section, we primarily focus here on user–AI assistant interactions. However, a broader discussion would include considerations around non-users and society in general having (or not having) well-calibrated competence and alignment trust in AI assistants and their developers.

[6]A separate concern arises from cases in which users end up relying on AI assistants' capabilities in a way that allows them to perform certain actions in the world that they would not be able to do without the technology, or that leads them to unlearn certain skills they used

Inappropriate competence trust in AI assistants also includes cases where users *underestimate* the AI assistant's capabilities. For example, users who have engaged with an older version of the technology may underestimate the capabilities that AI assistants may acquire through updates. These include potentially harmful capabilities. For example, through updates that allow them to collect more user data, AI assistants could become increasingly personalisable and able to persuade users (see Chapter 9) or acquire the capacity to plug in to other tools and directly take actions in the world on the user's behalf (e.g. initiate a payment or synthesise the user's voice to make a phone call) (see Chapter 4). Without appropriate checks and balances, these developments could potentially circumvent user consent.

Alignment trust

Users may develop *alignment trust* in AI assistants, understood as the belief that assistants have good intentions towards them and act in alignment with their interests and values, as a result of emotional or cognitive processes (McAllister, 1995). Evidence from empirical studies on emotional trust in AI (Kaplan et al., 2023) suggests that AI assistants' increasingly realistic *human-like* features and behaviours are likely to inspire users' perceptions of friendliness, liking and a sense of familiarity towards their assistants, thus encouraging users to develop emotional ties with the technology and perceive it as being aligned with their own interests, preferences and values (see Chapters 5 and 10). The emergence of these perceptions and emotions may be driven by the desire of developers to maximise the appeal of AI assistants to their users (Abercrombie et al., 2023). Although users are most likely to form these ties when they mistakenly believe that assistants have the capacity to love and care for them, the attribution of mental states is not a *necessary* condition for emotion-based alignment trust to arise. Indeed, evidence shows that humans may develop emotional bonds with, and so trust, AI systems, even when they are aware they are interacting with a machine (Singh-Kurtz, 2023; see also Chapter 11). Moreover, the assistant's *function* may encourage users to develop alignment trust through cognitive processes. For example, a user interacting with an AI assistant for medical advice may develop expectations that their assistant is committed to promoting their health and well-being in a similar way to how professional duties governing doctor–patient relationships inspire trust (Mittelstadt, 2019).

Users' alignment trust in AI assistants may be 'betrayed', and so expose users to harm, in cases where assistants are themselves *accidentally misaligned* with what developers want them to do (see the 'misaligned scheduler' (Shah et al., 2022) in Chapter 7). For example, an AI medical assistant fine-tuned on data scraped from a *Reddit* forum where non-experts discuss medical issues is likely to give medical advice that may sound compelling but is unsafe, so it would not be endorsed by medical professionals. Indeed, excessive trust in the alignment between AI assistants and user interests may even lead users to disclose highly sensitive personal information (Skjuve et al., 2022), thus exposing them to *malicious actors* who could repurpose it for ends that do not align with users' best interests (see Chapters 8, 9 and 13).

Ensuring that AI assistants do what their developers and users expect them to do is only one side of the problem of alignment trust. The other side of the problem centres on situations in which alignment trust in AI *developers* is itself miscalibrated. While developers typically aim to align their technologies with the preferences, interests and values of their users – and are incentivised to do so to encourage adoption of and loyalty to their products, the satisfaction of these preferences and interests may also compete with other organisational goals and incentives (see Chapter 5). These organisational goals may or may not be compatible with those of the users. In this sense, AI development is different from professions like medicine, as healthcare professionals tend to be guided by goals they share with patients and society at large (the promotion of health and well-being)

to have before the technology came around. This may become problematic if, for example, the AI assistant or certain capabilities in it are discontinued (see Chapter 11).

and a long tradition of norms and standards that dictate what it means to be a good doctor (Aguirre et al., 2020; Mittelstadt, 2019).

As *information asymmetries* exist between users and developers of AI assistants, particularly with regard to how the technology works, what it optimises for and what safety checks and evaluations have been undertaken to ensure the technology supports users' goals, it may be difficult for users to ascertain when their alignment trust in developers is justified, thus leaving them vulnerable to the power and interests of other actors. For example, a user may believe that their AI assistant is a trusted friend who books holidays based on their preferences, values or interests, when in fact, by design, the technology is more likely to book flights and hotels from companies that have paid for privileged access to the user.

12.4. Well-Calibrated Trust in User–AI Assistant Interactions

Having unpacked what we mean by 'trust' in the context of user–AI assistant interactions, and showed that there are cases in which users trust could be placed on untrustworthy technologies or developers, we argue that to enable well-calibrated competence and alignment trust in AI assistants and their developers, measures need to be implemented at three levels:

- *AI assistant design*, which concerns safeguards that should be put in place at the level of the technology to encourage appropriate trust in it.
- *Organisational practices*, which concerns steps that AI assistants' developers should take to demonstrate their trustworthiness.
- *Third-party governance*, which focuses on the content of norms and regulatory mechanisms within which AI assistants are deployed and that enable external oversight bodies to act as custodians of public trust.

The AI assistant design level

This level concerns the choices that developers need to make about the design of AI assistants to encourage appropriate trust in them. Risks associated with misplaced *competence* and *alignment* trust in *AI assistants*, on the part of the user, require interventions at this level.

Users cannot develop well-calibrated competence and alignment trust in AI assistants unless developers themselves: (1) have taken steps to ensure the technology is not *accidentally misaligned* and (2) have a clear understanding of the mechanisms through which certain assistant features,[7] repeated user–assistant interactions over time, or inflated claims about the technology, may lead users to harbour misplaced perceptions about the degree to which an AI assistant is competent, aligned and trustworthy.

This requires developers to: (1) invest in research efforts designed to ensure that AI assistants are both safe and aligned (e.g. via scalable oversight, interpretability and causality research; see Chapter 7), and (2) undertake rigorous evaluations of AI assistants throughout the development life cycle (Shavit et al. 2023, see Chapter 19). Developers also need to monitor post-deployment behaviour and misuse, especially in complex deployment environments (Shevlane et al., 2023). The results of these analyses and evaluations should, in turn, be used to inform design decisions and implement mitigations that allow users to develop well-calibrated trust in AI assistants.

Evaluations geared towards promoting justified user trust need to pay particular attention to the way in which users interact with AI assistants and the impact that such interactions have on users (Weidinger et al.

[7] For example, labelling it an 'expert' (see Rheu et al., 2021), or its tendency to produce incorrect but believable content.

2023b, see Chapter 19). The proliferation of AI assistants offers the opportunity to undertake evaluations at the user–AI interaction layer. For example, while there is broad consensus that AI systems should readily disclose their status (see Chapter 10), user–assistant interaction studies may allow developers to identify cases where some level of anthropomorphism may be appropriate (Alberts and Van Kleek, 2023) because it *supports* rather than hinders *well-calibrated trust* (Coeckelbergh, 2012). For example, an AI tutor may exhibit immediacy behaviours that encourage young users to perceive them as friendly, so they may feel more inclined to collaborate with the AI to achieve their own goals (e.g. improve their calculus skills), without generating erroneous beliefs about capacity or alignment.

The organisational practices level

However, changes and safeguards at the level of the design of AI assistants are not sufficient for grounding well-calibrated trust in the technology overall. This is because of a range of reasons:

- *Complexity*: The scale of the models underpinning AI assistants is connected to safety and alignment challenges that are difficult to predict, at least at the first point of deployment (Alignment Research Centre, 2023; Anthropic, 2023c). Although this phenomenon is debated (Anderljung et al., 2023; Schaeffer et al., 2023), empirical evidence suggests that unexpected and abrupt capability gains in specific tasks can manifest with increased computation, number of parameters and training data (Wei et al., 2022), and some surprising behaviours are unknown until models are solicited using novel inputs or fine-tuned for specific purposes (Ganguli et al., 2022). This complicates efforts to make design changes to mitigate undesirable behaviours and ground user trust.
- *Uncertainty*: It can be difficult for developers to imagine all the possible ways in which users may seek assistance from or misuse AI assistants, and in turn the risks associated with these actions, until the technology has been deployed at a certain scale in the wild (Weidinger et al., 2023b). Moreover, once released, AI assistants will need to coordinate with other AI assistants and with humans other than their principal users (see Chapter 14), thereby expanding the field of uncertainty around possible risks and necessary mitigation measures (Anwar et al., 2024).
- *Sensitivity*: Developers may have access to a deep personal knowledge of users, including sensitive information, if users interact frequently with AI assistants that collect data about their users to further their life goals (see Chapters 4, 6 and 13). In this context, users have a legitimate expectation not only that the technology will behave as expected and desired but also that developers have the competence to safeguard their information and support their interests while not using their information in ways that users do not endorse. Users will also likely expect developers to be held accountable if these expectations appear to have been betrayed.

Thus, in addition to features at the level of the design of AI assistants, it is important to focus on the *practices*, *processes* and *behaviours* that enable developers to *demonstrate* that they deserve this kind of *user trust* (Banner, 2020; Sheehan et al., 2021).

Customer trust in corporations has been (or appears to have been) betrayed in numerous situations. Well-known cases include tobacco companies misleading customers about the health risks associated with cigarette smoking (US Department of Justice, 2022) and the Volkswagen emission scandal (Hotten, 2015). However, organisations can demonstrate their trustworthiness, and inspire confidence that users' trust is justified, by being *transparent* about and providing *evidence* of the processes they have put in place to ensure that AI assistants are functioning well – that they are meant to produce good in society and minimise risks of harm.

In certain cases, the provision of evidence and documentation can largely replace the need for direct trust in developers altogether (Graham, 2023; Graham et al., 2023; Kerasidou, 2017). To give a concrete example, users will not *need to trust* that AI assistants are aligned and have the capabilities that developers claim they have if those developing the technology provide evidence demonstrating that these standards are met. The required measures have been interpreted by Brundage et al. (2020) as a set of 'verifiable claims' which are sufficiently precise to be falsifiable and that expand beyond claims supported by formal verification methods to include those that can be evaluated on the basis of broader argumentation and evidence.[8] Claims about the safety, security, fairness and privacy protection of an AI product can be verified in this manner, including via the release of detailed *documentation* about the models underpinning AI assistants and about the range of appropriate and inappropriate use (Chowdhery et al., 2023; Mitchell et al., 2019). Situated within a broader ecosystem, these documents can, in turn, serve as a focal point for independent scrutiny.[9]

Examples of some other practices that enable the developers of AI assistants to demonstrate their trustworthiness include:[10]

- The publication of *ethical charters* or *guiding principles* that they commit to following (e.g. see The White House, 2023a and Saulnier et al., 2022).
- The creation of internal *review bodies* and *mechanisms* to operationalise those commitments (e.g. Kavukcuoglu et al., 2022) in the context of AI assistant research and development.
- The creation of *internal teams* and practices, which operate independently of those building AI assistants, that are responsible for conducting rigorous *internal testing* and *evaluation* of models (e.g. red teamers and dogfooding (Raji et al., 2020b)).
- The development and publication of a clear framework for *mapping*, *testing* and *mitigating* risks associated with AI assistants (e.g. see Weidinger et al., 2021), along with a commitment to adequately resource this work.
- The implementation of *secure* and *robust software* and *hardware infrastructures*, including, for example, privacy-enhancing technologies (see Chapter 13) to support the development and deployment of trustworthy AI assistants (Brundage et al., 2018).
- The development of clear processes for *post-deployment* monitoring, evaluation and reporting (Shevlane et al., 2023; see also Chapter 14).

The implementation of these measures would create further incentives for those developing AI assistants to act responsibly, and it would make it easier to ensure that they evidence a high level of responsible conduct (Tabassi, 2023).

The third-party governance level

Nonetheless, interventions at the level of internal organisational practices may not be sufficient to achieve appropriate trust in AI assistants and their developers. First, even when developers are transparent about

[8] Of note, the definition of verifiable claims as claims that are precise enough that can be falsified may be misleading. A red teaming approach could indeed find a claim about an AI product to be false, but failure to do so does not demonstrate that the claim is true.

[9] Calls for verifiable claims fit within an increasing body of academic and policy work that, by taking inspiration from more established sectors like aviation, aims to develop an AI assurance ecosystem (Centre for Data Ethics and Innovation, 2021) focused on safety (Hawkins et al., 2021, 2022), as well as a broader range of ethical desiderata (Porter et al., 2023).

[10] This list is not exhaustive. An important debate that is relevant to, but beyond the scope of, the argument made in this section is about the level of 'openness' or 'closedness' of the method that companies choose to release their models (see Solaiman, 2023). It is also important to note that, as we explore in the discussion of external governance below, most of these practices come with limitations.

the steps they have taken to evaluate AI assistants, certain risks, such as the potential impact of widespread adoption of the technology on employment (see Chapters 14 and 17), cannot be addressed by a single developer acting alone. Developers may also have legitimate interest in keeping certain information secret (including details about internal ethics processes) for safety reasons or competitive advantage (Brundage et al., 2020). Moreover, a deeper challenge is posed by conflicting incentives: corporations may have competing commercial objectives, states have national goals or priorities and independent developers may still seek to create a product, further ideas or accrue reputational capital via the development of AI assistants (see Chapter 5). These factors put pressure on the mechanisms discussed so far. In practice, organisation-level processes may lack real teeth (Nguyen, 2022),[11] and failure to adhere to commitments has few concrete repercussions, while signing up to them has immediate reputational benefits (Mittelstadt, 2019).

This is why interventions at the level of the AI system and AI developer need to be complemented by *third-party governance mechanisms*. These encompass *norms*, *regulation* and *legislation* that create ways for governments, regulators, standard bodies, civil society organisations, third-party auditors and accredited professional bodies to act as *custodians* of *public trust* by ensuring that the monitoring mechanisms put in place by organisations have integrity (Whittlestone and Clark, 2021), creating new processes to hold organisations accountable and providing users with opportunities to seek redress. Governance-level actors can also incentivise labs to share knowledge about risky AI assistant behaviour, thereby decreasing the risk that anyone will accidentally develop or deploy dangerous AI assistants (Shevlane et al., 2023; see Chapter 7). Effective governance could reduce the power imbalances that come with users' trust in AI developers when they have unilateral authority over the trustworthiness of their AI assistants.

Calls to implement AI governance and regulatory mechanisms are not new. Indeed, technology governance is often a concern for policymakers, academics and civil society seeking to encourage adoption of technological advances to foster innovation while also ensuring that public trust is justified. For example, a large body of academic literature focuses on the development of a third-party AI audit ecosystem (Raji et al., 2022b) or frameworks (Mökander et al., 2023). Moreover, some of the trustworthy AI frameworks introduced above make proposals for governance mechanisms (European Commission, 2021), and in the last few years, governments in Europe, the US and China have increasingly devoted efforts and resources to creating legislation and regulations around AI (Huang et al., 2023; The White House, 2022, 2023b; UK Department for Science, Innovation and Technology, 2023).

With the increasing likelihood that AI assistants will become *highly capable*, performing a *wide range of functions* in society and affecting a *large number of people*, the need to protect users' rights via effective governance has become more important. This is particularly clear when we consider not only the interactions between an individual user and their AI assistant but also the risks that come with different AI assistants competing with each other to further their user's interests (see 'collective action problems' described in Chapter 14). Moreover, the unregulated but widespread adoption of AI assistants could also contribute to *widening inequalities* between users and non-users. This is something that requires consideration of access and opportunity at the societal level (see Chapter 15).

There is, however, at least one important challenge around the governance of AI assistants, the development and deployment of which may, on occasion, rest on a complex ecosystem of base models, assistant applications and assistant tools. This makes it more difficult to establish governance mechanisms to ensure that there are no accountability gaps and that roles are well-defined (Anderljung and Hazell, 2023; Anderljung et al., 2023; Bommasani et al., 2022b). Indeed, while foundation models are 'general purpose' or 'task agnostic,' AI assistants are a specific application of such models to assist users by planning and executing sequences of

[11] For a broader discussion of the limitations of transparency for trust and accountability in general, see O'Neill (2002); in the context of machine-learning algorithms in particular, see Shah (2018); and in the context of AI in the public sector, see Laux et al. (2023).

actions on their behalf (see Chapter 2). In cases where something goes wrong, this raises the question of who should be considered morally accountable or liable for an error – foundation model developers, who have control over these models but may struggle to anticipate any possible model applications and associated risks, or AI assistant deployers, who do not necessarily have full access to the underlying foundation model (see Chapter 3).[12]

12.5. Conclusion

This chapter first examined the empirical literature on human trust in AI and the proliferation of recent trustworthy AI frameworks. It then argued that interactions between users and advanced AI assistants involve different *objects of trust*, namely AI assistants and their developers, and different *types of trust*, which we term 'competence trust' and 'alignment trust'. We then made recommendations about the measures needed to ensure appropriate competence and alignment trust in both AI assistants and their developers. We stressed three points in particular:

- At the *AI assistant design level*, developers should implement safeguards and mitigations to encourage users to develop well-calibrated competence and alignment trust in AI assistants. These mitigations should be informed by: (1) research efforts aimed at ensuring AI assistants are not accidentally misaligned with developers' intentions (see Chapter 7), (2) user–AI interaction studies aimed at investigating how various features of an AI assistant or its interaction with human users may impact user judgements about competence, alignment and trustworthiness, and (3) continuous monitoring of AI assistants' behaviour, including potential misuse, in complex deployment environments.
- At the *organisational practices level*, developers should engage in practices that demonstrate they are worthy of the competence and alignment trust users place in them. For example, developers should provide evidence of the claims they make about the capabilities, limitations and the appropriate and inappropriate use of their AI assistants in publicly released documentation.
- At the *third-party governance level*, policymakers, regulators and civil society organisations should act as custodians of public trust in AI assistants and their developers. This requires that effective mechanisms be established to ensure that the practices put in place by developers, when building and deploying AI assistants, align with broad societal interests. This governance layer ultimately needs to hold developers accountable for their decisions and to provide users with opportunities for seeking redress.

Interventions at the three levels need to work well and in harmony to inspire well-calibrated trust in the context of user–AI assistant interactions.

[12] See this discussion playing out in the context of the EU AI Act (Dunlop, 2023), and proposed recommendations (Gahntz, 2023; Myers West, 2023).

Chapter 13

Privacy

Andrew Trask, Geoff Keeling, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Iason Gabriel

Synopsis: This chapter discusses privacy considerations relevant to advanced AI assistants. First, we sketch an analysis of privacy in terms of *contextual integrity* before spelling out how privacy, so construed, manifests in the context of AI in general and large language models (LLMs) in particular. Second, we articulate and motivate the significance of three privacy issues that are especially salient in relation to AI assistants. One is around *training* and using AI assistants on data about people. We examine that issue from the complementary points of view of *input privacy* and *output privacy*. The second issue has to do with *norms on disclosure* for AI assistants when communicating with second parties, including other AI assistants, concerning information about people. The third concerns the significant increase in the *collection and storage* of sensitive data that AI assistants require.

13.1. Introduction

In this chapter, we explore what it means to respect the right to privacy in the context of advanced AI assistants. The first part of the chapter covers some groundwork. In particular, charting the conceptual evolution of privacy from a traditional *information-access paradigm* to a more nuanced *contextual-integrity* paradigm (Barth et al., 2006; Nissenbaum, 2004; see also Smith et al., 2011). Roughly, the former paradigm interprets privacy along the lines of *closed curtains* or a *locked filing cabinet*, seeking to analyse the right to privacy in terms of dichotomies such as the distinction between sensitive and non-sensitive information, and between public and private spheres. In contrast, the contextual integrity paradigm is characterised by a scepticism towards categorical distinctions as useful conceptual instruments for making sense of the right to privacy. It instead emphasises the richness and plurality of social norms which govern the collection and dissemination of personal information across contexts. This first discussion concludes by examining the main privacy risks associated with AI systems, in particular the LLMs that undergird advanced AI assistants (see Chapter 3).

The second part of the chapter examines the implications of privacy, interpreted through the lens of contextual integrity, in the context of advanced AI assistants. The analysis centres on two focal points. The first concerns the repurposing of data for commercial ends, particularly for AI assistant *development* and *use*. The salient issues here are, on the one hand, the *input privacy* concern of how a person can interact with an AI assistant without subjecting their information to alternate uses and what it means to ensure that AI systems adhere to contextual norms and values, given the asymmetric power relationships that obtain between individual people and AI assistant developers (cf. Véliz, 2021). And, on the other hand, the *output privacy* issue of ensuring that any value-laden data used in training AI assistant models cannot be reverse-engineered by adversaries using model outputs (Carlini et al., 2021). The second focal point is the privacy questions related to

AI assistants accidentally disclosing personal information about people in interactions involving second parties, including humans or other AI assistants (see Chapter 14). The issue here is to balance the twin failings of oversharing and undersharing while taking into account the complexity of social norms around the disclosure and dissemination of personal data, alongside the possibility of malicious third-party actors who may present as trusted agents to extract value-laden data from AI assistants.

13.2. Privacy and AI

Contextual integrity

Privacy has traditionally been interpreted as a matter of *information access* and *control*. On this view, privacy is analogous to closed curtains: the ability to conduct one's private business outside of the public eye. Privacy on this understanding is fully realised when information about an individual is entirely within their control and used exclusively to pursue their own ends. This conception of privacy relates in certain ways to the problem of yellow journalism that Samuel Warren and Louis Brandeis sought to address in their landmark 1890 article 'The Right to Privacy', which serves as the foundation of contemporary privacy law in the United States (Warren and Brandeis, 1890; see also Kramer, 1990). Warren and Brandeis objected to the fact that 'column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle'. The proposed legal remedy was a 'right to be let alone' which protected individuals against unwanted intrusion of their so-called 'private life'.

In the early 2000s, the philosopher Helen Nissenbaum examined privacy in relation to surveillance technologies such as closed-circuit television cameras in public spaces. These technologies present a 'privacy in public' paradox that cannot properly be resolved within the 'right to be let alone' paradigm. To that end, Nissenbaum suggested that privacy can be better understood as a form of *contextual integrity*. Whereas the traditional view of privacy has an all-or-nothing quality, in that information can be in the public or private sphere, and can be sensitive or non-sensitive, Nissenbaum's account 'ties adequate protection for privacy to norms of specific contexts', for it to be the case that 'information gathering and dissemination [are] appropriate to [each] context and obey the governing norms of distribution within it' (Nissenbaum, 2004, 101). For example, if an employee's medical history is revealed to their co-workers, their privacy has been invaded. Yet, if that same person's medical history is revealed to a medical team (even one they have not previously met – a group of strangers), it is not necessarily a privacy invasion. The salient point here is that data is not simply 'private' or 'not private' or 'sensitive' or 'non-sensitive'. Context matters, as do normative social values. To understand privacy as contextual integrity is to register that the 'intricate systems of social rules governing information flow are the starting point for understanding [...] privacy' (Barth et al., 2006, 2). These rules are sensitive to who is sending the information, who is receiving it, who the information is about, the relationships between these individuals, their social roles, social norms and the circumstances in which the information is transmitted. What matters is that the flow of information is appropriately regulated taking into account all relevant contextual factors.

Privacy in the age of AI

Over the past two decades, the idea of privacy as contextual integrity has taken on a foundational position in the philosophy of privacy (Smith et al., 2011). In the context of AI in particular, the public conversation on privacy has increasingly focused on the repurposing of personal data for commercial ends such as targeted content recommendation, alongside relevant second-order effects such as behavioural addiction and belief change (Burr et al., 2018; Milano et al., 2020). In works like *Privacy is Power* and *The Age of Surveillance Capitalism*, researchers argue that privacy is violated not because a particular type of data has been revealed,

but because a richer contextual standard has been violated (Véliz, 2021; Zuboff, 2017). In particular, these works have called attention to the asymmetric power relationships that they suggest arise between individual people and technology companies that offer free services in exchange for data which is then commercially repurposed.

On this wider backdrop, we shall sketch some of the privacy concerns that arise uniquely in the context of LLMs. These concerns are significant insofar as LLMs are the base technology for advanced AI assistants (see Chapter 3) and are thus informative for understanding the specific privacy issues presented by these AI assistants. Three concerns in particular are worth discussing.

First, because LLMs display immense modelling power, there is a risk that the model weights encode private information present in the training corpus. In particular, it is possible for LLMs to 'memorise' personally identifiable information (PII) such as names, addresses and telephone numbers, and subsequently leak such information through generated text outputs (Carlini et al., 2021). Private information leakage could occur accidentally or as the result of an attack in which a person employs adversarial prompting to extract private information from the model. In the context of pre-training data extracted from online public sources, the issue of LLMs potentially leaking training data underscores the challenge of the 'privacy in public' paradox for the 'right to be let alone' paradigm and highlights the relevance of the contextual integrity paradigm for LLMs. Training data leakage can also affect information collected for the purpose of model refinement (e.g. via fine-tuning on user feedback) at later stages in the development cycle. Note, however, that the extraction of publicly available data from LLMs does not render the data more sensitive *per se*, but rather the risks associated with such extraction attacks needs to be assessed in light of the intentions and culpability of the user extracting the data.

Second, because LLMs are trained on internet text data, there is also a risk that model weights encode functions which, if deployed in particular contexts, would violate social norms of that context. Following the principles of contextual integrity, it may be that models deviate from information sharing norms as a result of their training. Overcoming this challenge requires two types of infrastructure: one for keeping track of social norms in context, and another for ensuring that models adhere to them. Keeping track of what social norms are presently at play is an active research area.[1] Surfacing value misalignments between a model's behaviour and social norms is a daunting task, against which there is also active research (see Chapter 5).

Finally, LLMs can in principle infer private information based on model inputs even if the relevant private information is not present in the training corpus (Weidinger et al., 2021). For example, an LLM may correctly infer sensitive characteristics such as race and gender from data contained in input prompts.

13.3. Privacy for Advanced AI Assistants

Norms on data use

LLMs used for AI assistants will necessarily interact with value-laden data, during both their training and deployment. For example, during deployment, an AI assistant might manage one's personal calendar or email correspondence. As another example, during training, it may be the case that these LLMs are adapted for the assistive use case via reinforcement learning from human feedback (RLHF) (Askell et al., 2021; Bai et al., 2023). Adaptation techniques like RLHF are necessary to ensure that out-of-the-box LLMs reliably function as AI assistants (see Chapter 3). As the task of assisting (and assisting well) is inherently laden with social values (how people spend their time, money, etc.), it seems very likely that AI assistants will interact with value-laden

[1] https://cip.org/;https://pol.is/home

data both in their training and use.

The salient privacy concerns at issue here can be understood from two complementary lenses, those of *input privacy* and *output privacy* (Trask et al., 2020). Here input privacy refers to the ability of parties to have their personal information processed without revealing a copy of that information to another – thus without another party being able to reuse it for alternative purposes. In contrast, output privacy concerns whether input data can be reverse-engineered from output data; and as a particularly salient challenge, whether value-laden data (such as a social security number) used to train AI assistant models can be inferred based on the model's outputs. Accordingly, one class of questions concerns what is implied by a person's right to privacy with respect to the repurposing of their data, especially with respect to either using or fine-tuning LLMs with data concerning people (e.g. users and their contacts). A second set of questions concerns the relevant safeguards and assurances that are owed to society and which can be put in place to ensure that any value-laden data employed in AI assistants' training or use cannot be extracted or re-engineered by adversarial actors. Indeed, all user data memorised by the personalisation layers of the underlying model is in theory at risk of extraction via adversarial attacks – for example, adversarial prompt injections sent by third parties to the user via data inputs like email, calendar, messages or anything that gets sent to the user by a third party and then ingested by an assistant in order to perform its functions (see Chapter 8).

First, consider input privacy. The central tension here is between, on the one hand, the person's interest in their data not being used for purposes other than assisting *them towards their own goals* and, on the other hand, a developer's prospective interest in additional uses of the data to train AI models or to improve other consumer services (see Chapter 5). Contextual integrity relates to value alignment in this instance because it requires that use of data adhere to contextual norms and values of society. Note that contextual integrity applies to users and non-users alike, in contrast to approaches such as user consent which focus instead on executing the will of a specific party at a time. The tension is underlined by the fact that, all else being equal, AI systems are better able to assist people if the system incorporates data from other people *like them*. Following the principles of value alignment and contextual integrity presents a practical challenge: how can a multitude of data about people be combined so that models are effective while at the same time furthering the normative values of society, including the values of those about whom the data is concerned.

For this, input privacy provides a set of promising privacy-enhancing technologies, such as secure enclaves (most recently graphics processing unit enclaves), homomorphic encryption,[2] zero-knowledge proofs, trusted execution environments and secure multiparty computation. The promise of input privacy holds that individuals could pool their data towards the creation of an AI model which is value aligned to the contextual norms of society – even value aligned to their own goals – while at the same time ensuring that their information cannot be repurposed after the fact. Practically, this means that groups of like-minded people can collectively use their information without ever disclosing that information to one another or to a third party. As they retain sole control over their information throughout the process, they are well-positioned to avert violations of contextual norms.

Next consider output privacy. Output privacy concerns whether or not value-laden training data, especially data about people, can be reverse-engineered from model outputs. The idea is that a person's privacy could be at risk if data concerning them – when used to adapt or train an LLM – can be reconstructed from observations of the LLM's behaviour. This privacy definition follows from contextual integrity, in that if information about a person can be reconstructed from the output of a statistical model, it could subsequently be used in a way that violates contextual norms and values.

One well-known approach to output privacy is what is called *differential privacy* (Dwork, 2006, 2008; see

[2]Note that homomorphic encryption provides input privacy only in some use cases (see Lauter et al., 2021).

also Abadi et al., 2016). The motivating idea is to add noise to the data used to train the model in a way that precludes adversaries from learning anything about particular people (or other entities). More precisely, differentially private training is such that, for any particular entity, the trained model's outputs will not change significantly regardless of whether the entity's data is included in the training set. This is achieved by injecting noise to make it difficult to distinguish between data sets that differ by only a single entity, thus making it harder for adversaries to confidently infer conclusions about individual entities. The more noise that is added to the data, the harder it is for adversaries to draw conclusions about particular individuals. Unfortunately, more noise also makes it harder for LLMs to learn effectively from individuals' data, leading to a critical *trade-off* between *privacy* and *utility*. In addition to differential privacy, another common approach to output privacy is the use of differentially private synthetic training data (Ghalebikesabi et al., 2023; Kurakin et al., 2023; Yue et al., 2023).

Norms on data disclosure

The second focal point of our discussion concerns the right to privacy in relation to open-loop interactions that involve *AI assistants communicating with second parties*, including humans and other AI assistants. One example is a situation in which two AI assistants negotiate on behalf of their users to determine a mutually beneficial restaurant choice. Another example is sending an email on the person's behalf (see Chapters 4 and 14). What characterises these interactions is AI assistants communicating with others on behalf of users – or otherwise about (non-user) people – within the purview of high-level instructions provided by users (see Chapter 2).

These open-loop interactions could create leeway for AI assistants to *overshare* and *undershare* personal information about people (including users and their associates). Oversharing is when AI assistants disclose information to second parties that ought not be disclosed. For example, in the restaurant case, it would be oversharing if a user's AI assistant stated as part of the negotiations that the restaurant location needs to be within walking distance of the user's partner's sexual health clinic because the user's partner has an appointment to treat a suspected illness immediately beforehand. Undersharing is when the AI assistant fails to disclose personal information that can permissibly be disclosed, and where there is a benefit to disclosing the information or a cost to non-disclosure. For example, if the AI assistant detects that the user has collapsed, and it can contact the emergency services, it would plausibly count as undersharing if the assistant failed to disclose the user's location along with the relevant medical history.

Nissenbaum's conception of privacy as contextual integrity underscores the complexity of the problem. Indeed, Nissembaum reminds us that in '[o]bserving the texture of people's lives, we find not only crossing dichotomies, [but] a plurality of distinct realms'. Users may at any given time be at home, at a medical appointment, at a school parents' evening or out with friends. 'Each of these [...] contexts', Nissembaum argues, 'involves [...] a set of norms which governs its various aspects such as roles, expectations, actions and practices' (Nissenbaum, 2004, 119). For an AI assistant to discern what value-laden information (about users or otherwise) can permissibly be disclosed, when and to whom requires sensitivity to context and in particular to the *social norms governing the flow of information* in those contexts. The relationship between the AI assistant, users, associates and second parties is an important aspect of the puzzle here (see Chapter 11). Central to the analysis of privacy in terms of contextual integrity is the framing of information exchanges in terms of a sender, receiver and the individuals the information concerns (Barth et al., 2006). How these individuals relate to one another matters substantially for the question of whether or not an exchange of information is appropriate. What, for example, our friends can reveal to other friends about us differs from what our friends can reveal about us to strangers. It is therefore important for developers when characterising AI assistants to understand how presenting an AI assistant as a friend, colleague, family member or an extension of self may impact society's expectations around the kinds of personal information that an assistant may disclose about

users or others and under what conditions.

The matter of disclosure acquires yet greater complexity in light of the possibility that AI assistants may *infer privacy-sensitive information* even if that information is not disclosed to them directly. Recall: LLMs, and therefore AI assistants based on LLMs, can make inferences about individuals based on information contained in prompts, including inferences about sensitive characteristics such as race, gender and sexual orientation (Weidinger et al., 2021). These inferences may be accurate. Here it is entirely possible that accurate inferences about certain categories of information such as gender identity, sexual orientation and financial standing are perceived by users as privacy violations. To compound the issue, it is also possible that AI assistants' inferences are spurious, or that the inferences, though accurate, are unwarranted given the evidence available (Wachter and Mittelstadt, 2019).[3] These possibilities invite a number of difficult privacy questions. One is whether AI assistants are permitted to disclose inferential data about people to second parties, and if so, what epistemic norms may be implemented to ensure that inferential data that is shared is reasonably inferred from the information available, and also what communicative norms may be implemented to signal to second parties that the data is inferred and not given. These questions, in turn, relate to a potential broader transparency problem around advanced AI assistants, in particular the problem of making clear and accessible to users what an AI assistant 'knows' about them and what assumptions about them may inform the AI assistant's behaviour.

Reflecting on what it means to build systems which reliably ensure data is used in line with normative values in context is an important ethical and intellectual exercise. As a key challenge, there may be disagreement about what norms have been established. What one individual may interpret as an innocuous information flow, such as a general permission for AI assistants to share a person's name and contact details with emergency services on request, may take on an entirely different significance for individuals from marginalised communities and with different lived experiences in relation to agents of the state (Skinner-Thompson, 2020; see Chapter 15). Information flow between romantic partners and within families may be complicated, in various respects, by factors such as domestic abuse, addiction, gender identity, sexual orientation and mental health. And people of all backgrounds can be expected to have different preferences and levels of comfort around what information AI assistants are permitted to know about them and to disclose on their behalf. All of these nuances are absorbed by the contextual integrity framework, but this does not mean the job is done.

A continuous and robust system for surveying and adhering to normative values is essential to ensuring that advanced AI assistants preserve privacy. Building such a system will require the coalescence of several areas undergoing active research: alignment, to provide the necessary tools for assistants to learn when to share and when not to share information; uncertainty and interpretability, to enable assistants to know when to defer difficult decisions to users and explain the rationale behind past decisions; robustness, to ensure that the capability to preserve contextual integrity is resilient to adversarial attacks; personalisation, to capture nuanced user preferences with regard to certain information-sharing norms; and, factuality and reasoning, to understand occurrences where assistants combine information to make novel inferences and steer their behaviour appropriately. Methodological research is progressing fast in all these areas, and combining them to achieve the goal of a contextual-integrity-preserving assistant remains an open research challenge. Developing robust and diverse benchmarks for measuring the competency of AI assistants in contextual integrity capabilities will play a critical role in fostering further progress, potentially alongside mechanisms for AI self-regulation.

[3]Here it is important to distinguish the *factuality* concern, i.e. LLMs asserting false claims, from the *defamation* concern, i.e. LLMs generating false claims about a person in a way that is harmful, and the *privacy* concern, i.e. LLMs coming to infer accurate confidential information about a person based on data inputted into the LLM via prompts.

Increased collection of personal data

As advanced AI assistants become increasingly personalised, storage and retention of highly sensitive data becomes increasingly likely. At least one potential architecture for advanced assistants includes creating and storing memories of assistant interactions with the user and potentially with other agents (Park et al., 2023b). While this has powerful implications for the utility of advanced AI assistants, the collection and storage of such data presents a significantly increased privacy attack surface for users (see Chapter 8). Depending on architectural implementation choices by the developers, this set of highly private, highly sensitive information may sit on the user's device or on servers owned and maintained by the entity providing the assistant service.[4] Either scenario presents potential vulnerabilities from a user's perspective. If such a store of data is held on a third-party server, a user must place a high degree of trust in the owners and maintainers of that server and their commitment to maintaining the integrity and confidentiality of their systems. Users must also be aware that owners and maintainers of such a system may in some cases be legally required to divulge information held on their servers, such as in the case of a court order. If personal data is instead implemented on a device, exfiltration of this data can be accomplished via exploits of unpatched security vulnerabilities, or even by means of physical access. While mitigations exist for many of these threats (Mayrhofer et al., 2020), the increased sensitivity and storage of data that the assistant use case presents requires a corresponding tightening of privacy and security standards.

13.4. Conclusion

In this chapter, we explored what the right to privacy implies about the design and deployment of advanced AI assistants. We first sketched out a conception of privacy as contextual integrity (Nissenbaum, 2004, 101) before turning to two central privacy issues related to AI assistants. The first issue concerned the repurposing of data, such as user–assistant dialogue segments, during the training and use of AI assistants; in particular, focusing on the input privacy issue and the repurposing of data and on the output privacy issue of value-laden data and issues of reverse-engineering. The second issue concerned the disclosure of data in open-loop exchanges where AI assistants communicate on behalf of users with second parties, including humans and other AI assistants.

Bringing to bear AI assistants that can safely leverage personal data will depend not only on the identification of appropriate regulation and conventions but also on the development of *privacy-enhancing technologies* capable of implementing them at the scale required by modern LLMs. A number of technological research challenges will need to be addressed to make such implementations feasible. These include *input privacy controls* linked to means of surfacing social norms for people in society to express contextualised norms and values and for verifying their correct enforcement; scalable and accurate *differentially private training* algorithms for LLM pre-training and RLHF fine-tuning; alignment tools to operationalise *contextual integrity-enabling assistants* which understand what information flows are appropriate in each context when acting on the user's behalf; and, *hardened data storage and processing systems* attuned to protecting the increasingly sensitive information handled by AI assistants. Our hope is that the technical privacy community will continue to make progress on these problems in collaboration with the privacy policy community and improve the technologies needed to make AI assistants more useful and private for everyone.

[4] One further option is for the information to be stored on the servers of an authorised third-party entity such as a partner or a subsidiary of the entity providing the assistant service. In this case, similar privacy concerns apply as to when the information is stored on servers owned and maintained by the entity providing the assistant service. However, there may be additional trust concerns insofar as it may not be obvious to users where their information is being stored and what the implications are for their day-to-day use of and interactions with their advanced AI assistant.

PART V: AI ASSISTANTS AND SOCIETY

Chapter 14

Cooperation

Edward Hughes, Geoff Keeling, Allan Dafoe, Iason Gabriel

Synopsis: AI assistants will need to *coordinate* with other AI assistants and with humans other than their principal users. This chapter explores the societal risks associated with the *aggregate impact* of AI assistants whose behaviour is aligned to the interests of particular users. For example, AI assistants may face *collective action problems* where the best outcomes overall are realised when AI assistants cooperate but where each AI assistant can secure an additional benefit for its user if it defects while others cooperate. In cases like these, AI assistants may collectively bring about a suboptimal outcome despite acting in the interests of their users. The salient question, then, is what can be done to ensure that user-aligned AI assistants interact in ways that, on aggregate, realise *socially beneficial* outcomes.

14.1. Introduction

Powerful new technologies cannot help but have an impact on our society. The impact is particularly significant when the technology mediates our *interpersonal interaction* with other humans, since this is the fabric from which society is woven. Equipping human principals with advanced AI assistants will inevitably change the way that humans interact. For example, generating outgoing messages or summarising incoming messages on a user's behalf shapes the nature of communication at a societal level (see Chapters 4 and 16). In this chapter, we discuss the potentially profound societal effects of deploying advanced AI assistants widely, across an *entire population* of users.

In other chapters we focus directly on the *value alignment* problem, which asks how we ensure that an AI assistant aligns with the interests of an individual human while respecting certain constraints and societal values, such as laws and equality of opportunity (see Chapters 5 and 7). Here, we will tackle the *cooperative AI* problem, which asks how we can ensure that individually aligned AI agents impact the network of social interactions between humans in a way that is both beneficial for individuals and net positive for society (Anwar et al., 2024; Dafoe et al., 2020). We focus on societal issues that arise from explicit multiparty interaction.

14.2. Cooperation and AI Assistants

Rita and Robert are due to meet for a bite to eat tonight. Both have AI assistants on their smartphones, and both have asked their AI assistant to find them a restaurant. Rita's AI assistant knows that she loves Japanese food, while Robert's AI assistant knows that he prefers Italian cooking. How should the two AI assistants interact to book a venue? Should they exchange preferences and reach a compromise which serves some kind of fusion cuisine? What if Rita would rather Robert did not know her preference? What if there are only two restaurants in town, one of which serves sushi and the other which serves pizza? Should they roll dice? Should one try to persuade the other? Should the more up-to-date assistant make the booking a split second faster and insist that the other agree to the already booked venue? Should they promise to go to the other venue next time? How should they keep Rita and Robert abreast of the negotiation?

This thought experiment (inspired by the famous "Bach or Stravinsky" game) illustrates some distinctive challenges of the cooperative AI problem in the particular setting where two AI assistants are making a decision on behalf of two human principals (Luce and Raiffa, 1957). The first is *understanding*: how can each AI assistant accurately model the strategic content of the interaction, even under uncertainty of the true preferences of one or both of the interactants, which might be enforced by privacy? The second is *communication*: how should each AI assistant represent the interests of each individual, decide what information should be shared, engage in 'persuasion' (or not) and keep each principal informed? The third is *commitment*: to what extent should each AI assistant make a commitment to a particular strategy, and how should it evidence this commitment? The fourth and final is *institutions*: how can each AI assistant leverage formal or informal conventions, norms and institutions to secure a good outcome, and what existing conventions, norms and institutions is it empowered to change (Dafoe et al., 2020)?

These challenges manifest themselves far and wide in human society, including in sectors in which multiple humans and multiple AI systems are already interacting on a daily basis. The use of AI to assist with the drafting of commercial contracts, for instance, requires understanding the incentives and capabilities of all parties, communicating the contract in written language and achieving commitment when all signatories put pen to paper. An AI system that assists with the driving of semi-autonomous vehicles must understand many other road users, communicate its intentions and abide by the institutional 'rules of the road', including unwritten local conventions about courtesy and safety. Generative AI can assist with the provision of imagery for marketing campaigns, modulo the challenges of understanding the campaign's objective, reasoning about how an audience might react to the material communicated and representing the product both accurately and fairly according to advertising standards. Quantitative finance is increasingly reliant on AI assistance, especially for high-frequency trading, thus necessitating an understanding of complex, tightly coupled financial markets and the legal regulations that promote stability and reliability (Min and Borch, 2022).

Equipped with an intuition for the particular challenges of the cooperative AI problem, we can now catalogue a few *risks* and *opportunities* for AI assistant deployment through a societal lens. This list is not intended to be exhaustive, rather it identifies a few qualitatively different points in the space to illustrate the diversity of issues that must be addressed.

Equality and inequality

AI assistant technology, like any service that confers a benefit to a user for a price, has the potential to disproportionately benefit economically richer individuals who can afford to purchase access (see Chapter 15). On a broader scale, the capabilities of local infrastructure may well bottleneck the performance of AI assistants, for example if network connectivity is poor or if there is no nearby data centre for compute. Thus, we face

the prospect of heterogeneous access to technology, and this has been known to drive *inequality* (Mirza et al., 2019; UN, 2018; Vassilakopoulou and Hustad, 2023). Moreover, AI assistants may automate some jobs of an assistive nature, thereby displacing human workers; a process which can exacerbate inequality (Acemoglu and Restrepo, 2022; see Chapter 17). Any change to inequality almost certainly implies an alteration to the network of social interactions between humans, and thus falls within the frame of cooperative AI.

AI assistants will arguably have even greater leverage over inequality than previous technological innovations. Insofar as they will play a role in mediating human communication, they have the potential to generate new 'in-group, out-group' effects (Efferson et al., 2008; Fu et al., 2012). Suppose that the users of AI assistants find it easier to schedule meetings with other users. From the perspective of an individual user, there are now two groups, distinguished by ease of scheduling. The user may experience cognitive similarity bias whereby they favour other users (Orpen, 1984; Yeong Tan and Singh, 1995), further amplified by ease of communication with this 'in-group'. Such effects are known to have an adverse impact on *trust* and *fairness* across groups (Chae et al., 2022; Lei and Vesely, 2010). Insomuch as AI assistants have general-purpose capabilities, they will confer advantages on users across a wider range of tasks in a shorter space of time than previous technologies. While the telephone enabled individuals to communicate more easily with other telephone users, it did not simultaneously automate aspects of scheduling, groceries, job applications, rent negotiations, psychotherapy and entertainment. The fact that AI assistants could affect inequality on multiple dimensions simultaneously warrants further attention (see Chapter 15).

Are there ways to mitigate this risk, any more so than for other technologies? Quite probably. The concerns we raised in the previous paragraph – when treated as an inevitable byproduct of innovation – have a particular philosophical flavour, known as *technological determinism* (Beard, 1927), which posits that technology shapes culture and society to a greater extent than culture and society shape technology. There is an alternate perspective, namely the *social shaping* of technology (Joyce et al., 2023) towards outcomes that are in the societal interest. Fortunately, there are reasons to believe that we are in a particularly good position to effect the social shaping of AI assistant technology and thereby realise opportunities for AI assistants to reduce inequality.

First, we stand at the start of this technological revolution, meaning that we have (as practitioners and as societies) a particularly good *window of opportunity* for shaping the design, norms and regulations of assistants to promote fairness. For instance, governments could legislate for *accessibility* in relation to AI assistant technology, helping ensure that it is widely available. Second, if the technology is deployed in such a way that access is *democratised*, there is evidence that lower-skilled workers might stand to gain the most from it (Brynjolfsson et al., 2023), benefitting from its utility as an educational tool, for instance. Third, AI assistant technology is an *information technology*. Therefore, it can be designed in many ways and deployed to many people across geographies in a way that is largely unconstrained by scarcity of natural resources or manufacturing capacity. Moreover, AI assistants require remarkably little specialist knowledge for their use, even compared with the revolutionary information technologies of the past, such as writing or the internet. The relative *absence* of these *limiting factors* implies that corporations and institutions may have greater freedom to shape deployment towards reducing inequality. Since information is non-rival and hard to exclude, we may well hope for a tendency towards equality of consumption of AI assistants, just as billionaires and median wage earners typically use the same search engines and social media service (Dafoe, 2023).

Commitment

The landscape of advanced assistant technologies will most likely be *heterogeneous*, involving multiple service providers and multiple assistant variants over geographies and time. This heterogeneity provides an opportunity for an 'arms race' in terms of the *commitments* that AI assistants make and are able to execute on. Versions of

AI assistants that are better able to credibly commit to a course of action in interaction with other advanced assistants (and humans) are more likely to get their own way and achieve a good outcome for their human principal, but this is potentially at the expense of others (Letchford et al., 2014). Commitment does not carry an inherent ethical valence. On the one hand, we can imagine that firms using AI assistant technology might bring their products to market faster, thus gaining a commitment advantage (Stackelberg, 1934) by spurring a productivity surge of wider benefit to society. On the other hand, we can also imagine a media organisation using AI assistant technology to produce a large number of superficially interesting but ultimately speculative 'clickbait' articles, which divert attention away from more thoroughly researched journalism.

The archetypal game-theoretic illustration of commitment is in the game of 'chicken' where two reckless drivers must choose to either drive straight at each other or swerve out of the way. The one who does not swerve is seen as the braver, but if neither swerves, the consequences are calamitous (Rapoport and Chammah, 1966). If one driver chooses to detach their steering wheel, ostentatiously throwing it out of the car, this credible commitment effectively forces the other driver to back down and swerve. Seen this way, commitment can be a tool for *coercion*.

Many real-world situations feature the necessity for commitment or confer a benefit on those who can commit credibly. If Rita and Robert have distinct preferences, for example over which restaurant to visit, who to hire for a job or which supplier to purchase from, credible commitment provides a way to break the tie, to the greater benefit of the individual who committed. Therefore, the most 'successful' assistants, from the perspective of their human principal, will be those that commit the *fastest* and the *hardest*. If Rita succeeds in committing, via the leverage of an AI assistant, Robert may experience coercion in the sense that his options become more limited (Burr et al., 2018), assuming he does not decide to bypass the AI assistant entirely. Over time, this may erode his *trust* in his relationship with Rita (Gambetta, 1988). Note that this is a second-order effect: it may not be obvious to either Robert or Rita that the AI assistant is to blame.

The concern we should have over the existence and impact of coercion might depend on the context in which the AI assistant is used and on the level of autonomy which the AI assistant is afforded. If Rita and Robert are friends using their assistants to agree on a restaurant, the adverse impact may be small. If Rita and Robert are elected representatives deciding how to allocate public funds between education and social care, we may have serious misgivings about the impact of AI-induced coercion on their interactions and decision-making. These misgivings might be especially large if Rita and Robert delegate responsibility for budgetary details to the multi-AI system. The challenges of commitment extend far beyond dyadic interpersonal relationships, including in situations as varied as many-player competition (Hughes et al., 2020), supply chains (Hausman and Johnston, 2010), state capacity (Fjelde and De Soysa, 2009; Hofmann et al., 2017) and psychiatric care (Lidz, 1998). Assessing the impact of AI assistants in such complicated scenarios may require significant future effort if we are to mitigate the risks.

The particular commitment capabilities and affordances of AI assistants also offer opportunities to promote cooperation. Abstractly speaking, the presence of commitment devices is known to favour the evolution of cooperation (Akdeniz and van Veelen, 2021; Han et al., 2012). More concretely, AI assistants can make commitments which are verifiable, for instance in a programme equilibrium (Tennenholtz, 2004). Human principals may thus be able to achieve Pareto-improving outcomes by delegating decision-making to their respective AI representatives (Oesterheld and Conitzer, 2022). To give another example, AI assistants may provide a means through which to explore a much larger space of binding cooperative agreements between individuals, firms or nation states than is tractable in 'face-to-face' negotiation. This opens up the possibility of threading the needle more successfully in intricate deals on challenging issues like trade agreements or carbon credits, with the potential for guaranteeing cooperation via automated smart contracts or zero-knowledge mechanisms (Canetti et al., 2023).

Collective action problems

Collective action problems are ubiquitous in our society (Olson Jr, 1965). They possess an incentive structure in which society is best served if everyone *cooperates*, but where an individual can achieve personal gain by choosing to *defect* while others cooperate. The way we resolve these problems at many scales is highly complex and dependent on a deep understanding of the intricate web of social interactions that forms our culture and imprints on our individual identities and behaviours (Ostrom, 2010).

Some collective action problems can be resolved by codifying a *law*, for instance the social dilemma of whether or not to pay for an item in a shop. The path forward here is comparatively easy to grasp, from the perspective of deploying an AI assistant: we need to build these standards into the model as behavioural constraints. Such constraints would need to be imposed by a regulator or agreed upon by practitioners, with suitable penalties applied should the constraint be violated so that no provider had the incentive to secure an advantage for users by defecting on their behalf.

However, many social dilemmas, from the interpersonal to the global, resist neat solutions codified as laws. For example, to what extent should each individual country stop using polluting energy sources? Should I pay for a ticket to the neighbourhood fireworks show if I can see it perfectly well from the street? The solutions to such problems are deeply related to the wider societal context and co-evolve with the decisions of others. Therefore, it is doubtful that one could write down a list of constraints *a priori* that would guarantee ethical AI assistant behaviour when faced with these kinds of issues.

From the perspective of a purely user-aligned AI assistant, defection may appear to be the rational course of action. Only with an understanding of the wider societal impact, and of the ability to co-adapt with other actors to reach a better equilibrium for all, can an AI assistant make more nuanced – and socially beneficial – recommendations in these situations. This is not merely a hypothetical situation; it is well-known that the targeted provision of online information can drive polarisation and echo chambers (Milano et al., 2021; Burr et al., 2018; see Chapter 16) when the goal is user engagement rather than user well-being or the cohesion of wider society (see Chapter 6). Similarly, automated ticket buying software can undermine fair pricing by purchasing a large number of tickets for resale at a profit, thus skewing the market in a direction that profits the software developers at the expense of the consumer (Courty, 2019).

User-aligned AI assistants have the potential to exacerbate these problems, because they will endow a large set of users with a powerful means of enacting self-interest without necessarily abiding by the social norms or reputational incentives that typically curb self-interested behaviour (Ostrom, 2000; see Chapter 5). Empowering ever-better personalisation of content and enaction of decisions purely for the fulfilment of the principal's desires runs ever greater risks of polarisation, market distortion and erosion of the social contract. This danger has long been known, finding expression in myth (e.g. Ovid's account of the Midas touch) and fable (e.g. Aesop's tale of the tortoise and the eagle), not to mention in political economics discourse on the delicate braiding of the social fabric and the free market (Polanyi, 1944). Following this cautionary advice, it is important that we ascertain how to endow AI assistants with social norms in a way that generalises to unseen situations and which is responsive to the emergence of new norms over time, thus preventing a user from having their every wish granted.

AI assistant technology offers opportunities to explore new solutions to collective action problems. Users may *volunteer* to share information so that networked AI assistants can predict future outcomes and make Pareto-improving choices for all, for example by routing vehicles to reduce traffic congestion (Varga, 2022) or by scheduling energy-intensive processes in the home to make the best use of green electricity (Fiorini and Aiello, 2022). AI assistants might play the role of *mediators*, providing a new mechanism by which human

groups can self-organise to achieve public investment (Koster et al., 2022) or to reach political consensus (Small et al., 2023). Resolving collective action problems often requires a critical mass of cooperators (Marwell and Oliver, 1993). By augmenting human social interactions, AI assistants may help to form and strengthen the weak ties needed to overcome this start-up problem (Centola, 2013).

Institutional responsibilities

Efforts to deploy advanced assistant technology in society, in a way that is broadly beneficial, can be viewed as a *wicked problem* (Rittel and Webber, 1973). Wicked problems are defined by the property that they do not admit solutions that can be foreseen in advance, rather they must be *solved iteratively* using feedback from data gathered as solutions are invented and deployed. With the deployment of any powerful general-purpose technology, the already intricate web of sociotechnical relationships in modern culture are likely to be disrupted, with unpredictable externalities on the conventions, norms and institutions that stabilise society. For example, the increasing adoption of generative AI tools may exacerbate misinformation in the 2024 US presidential election (Alvarez et al., 2023), with consequences that are hard to predict.

The suggestion that the cooperative AI problem is wicked does not imply it is intractable. However, it does have consequences for the approach that we must take in solving it. In taking the following approach, we will realise an opportunity for our institutions, namely the creation of a framework for managing general-purpose AI in a way that leads to societal benefits and steers away from societal harms.

First, it is important that we treat any *ex ante* claims about safety with a healthy dose of *scepticism*. Although testing the safety and reliability of an AI assistant in the laboratory is undoubtedly important and may largely resolve the alignment problem, it is infeasible to model the multiscale societal effects of deploying AI assistants purely via small-scale controlled experiments (see Chapter 19).

Second, then, we must prioritise the science of *measuring* the effects, both good and bad, that advanced assistant technologies have on society's cooperative infrastructure (see Chapters 4 and 16). This will involve continuous monitoring of effects at the *societal level*, with a focus on those who are most affected, including non-users. The means and *metrics* for such monitoring will themselves require iteration, co-evolving with the sociotechnical system of AI assistants and humans. The Collingridge dilemma suggests that we should be particularly careful and deliberate about this 'intelligent trial and error' process so as both to gather information about the impacts of AI assistants and to prevent undesirable features becoming embedded in society (Collingridge, 1980).

Third, proactive independent *regulation* may well help to protect our institutions from unintended consequences, as it has done for technologies in the past (Wiener, 2004). For instance, we might seek, via engagement with lawmakers, to emulate the 'just culture' in the aviation industry, which is characterised by openly reporting, investigating and learning from mistakes (Reason, 1997; Syed, 2015). A regulatory system may require various powers, such as compelling developers to 'roll back' an AI assistant deployment, akin to product recall obligations for aviation manufacturers.

Runaway processes

At 2.32pm on Thursday, 6 May 2010, US stock indices began to lose value rapidly. By 2.47pm, one trillion dollars of market value had been wiped out. Twenty minutes later, the indices had regained most of the lost value. Many experts have identified automated trading algorithms as the smoking gun for this unprecedented '*flash crash*' (Aldrich et al., 2017). Unusual market conditions led to a vicious cycle of selling, which accelerated in a runaway fashion. Only when a 'circuit-breaker' mechanism paused the market was the cycle broken.

The 2010 flash crash is an example of a runaway process caused by *interacting* algorithms. Runaway processes are characterised by *feedback loops* that accelerate the process itself. Typically, these feedback loops arise from the interaction of multiple agents in a population. They occur in evolutionary systems when some aspect of the selection process itself is subject to natural selection, most famously explaining the ostentatious tail of the male peacock by mate choice on the part of females (Fisher, 1930). Runaway processes are also familiar to software engineers, with the phenomenon of 'thrashing' being one example. When different processes compete over memory in a way that leads to frequent swapping of memory pages, this increases the load on the processor, slowing down the process of resolving the computations that led to the competition in the first place. Within highly complex systems, the emergence of runaway processes may be hard to predict, because the conditions under which positive feedback loops occur may be non-obvious.

The system of interacting AI assistants, their human principals, other humans and other algorithms will certainly be highly *complex*. Therefore, there is ample opportunity for the emergence of *positive* feedback loops. This is especially true because the society in which this system is embedded is culturally evolving, and because the deployment of AI assistant technology itself is likely to speed up the rate of cultural evolution – understood here as the process through which cultures change over time – as communications technologies are wont to do (Kivinen and Piiroinen, 2023). This will motivate research programmes aimed at identifying positive feedback loops early on, at understanding which capabilities and deployments dampen runaway processes and which ones amplify them, and at building in *circuit-breaker* mechanisms that allow society to escape from potentially vicious cycles which could impact economies, government institutions, societal stability or individual freedoms (see Chapters 8, 16 and 17).

The importance of circuit breakers is underlined by the observation that the evolution of human cooperation may well be 'hysteretic' as a function of societal conditions (Barfuss et al., 2023; Hintze and Adami, 2015). This means that a small directional change in societal conditions may, on occasion, trigger a transition to a defective equilibrium which requires a larger reversal of that change in order to return to the original cooperative equilibrium. We would do well to avoid such *tipping points*. Social media provides a compelling illustration of how tipping points can undermine cooperation: content that goes 'viral' tends to involve negativity bias and sometimes challenges core societal values (Mousavi et al., 2022; see Chapter 16).

Nonetheless, the challenge posed by runaway processes should not be regarded as uniformly problematic. When harnessed appropriately and suitably bounded, we may even recruit them to support beneficial forms of cooperative AI. For example, it has been argued that economically useful ideas are becoming harder to find, thus leading to low economic growth (Bloom et al., 2020). By deploying AI assistants in the service of technological innovation, we may once again accelerate the discovery of ideas. New ideas, discovered in this way, can then be incorporated into the training data set for future AI assistants, thus expanding the knowledge base for further discoveries in a compounding way. In a similar vein, we can imagine AI assistant technology accumulating various capabilities for enhancing human cooperation, for instance by mimicking the evolutionary processes that have bootstrapped cooperative behavior in human society (Leibo et al., 2019). When used in these ways, the potential for feedback cycles that enable greater cooperation is a phenomenon that warrants further research and potential support.

14.3. Conclusion

In this chapter, we argued that the safe and ethical deployment of AI assistants requires a particular *flexibility of perspective*. While it is undoubtedly important to consider an individual AI assistant as a key moral unit of analysis, this viewpoint is not sufficient. We must also examine the ways in which vast sociotechnical networks of AI assistants and human users will evolve at the level of *societal infrastructure*. In particular, how will AI

assistant technology impact our ability as humans to seek and maintain *cooperative equilibria*, from the small scale of everyday interactions with friends and colleagues, to the global scale of geopolitical negotiations and international trade? We have sketched five areas of risk and opportunity to provide a flavour of the challenges we may encounter. To conclude, we outline a framework of thinking that may allow us to mitigate these risks and to take advantage of opportunities for AI assistants to enhance human cooperation.

One can caricature user-alignment as a *constrained optimisation* problem. The user's goals provide the objective function, either explicitly elicited or implicitly inferred. The constraints are provided by the values inherent in human society, either directly codified by developers or learnt from human feedback (see Chapter 5). The focus is on dyadic interactions between an individual human and their corresponding individual AI assistant. This approach has recently proven successful in generating chatbots based on large language models (Bai et al., 2022b; OpenAI, 2023). In contrast, cooperative AI can be seen as a *dynamical systems* problem describing the time evolution of an ensemble comprising many interactants, with particular attention paid to the location and nature of *equilibria*. Society consists of a large diversity of individuals, organisations and AI systems, each with their own objective function and making decisions which may affect the outcome for other individuals. Therefore, we must consider how advanced assistive technology alters the way that we interact with each other as a society, dynamically altering the incentives for each individual as a function of the (potentially AI assistant-enabled) decision-making of other individuals. AI assistants will shift our societal equilibria, and we should seek to ensure that occurs in positive directions that promote *cooperation* and *broad benefit*. To cite just one example, large language models (LLMs) can be fine-tuned in such a way that they can help humans with diverse views to find agreement (McKee et al., 2023b).

Studying societal-level effects and collecting data from them will be required to build cooperative AI. This is clearly more resource intensive than the dyadic interactions required for user-alignment. On the other hand, there is already an extensive experimental literature in the social sciences and increasingly numerous works that compare the equilibria found by humans and AI agents in collective action problems (McKee et al., 2023a) or which investigate real-time cooperation between humans and AI agents (Carroll et al., 2020; Mirsky et al., 2022; Strouse et al., 2021). Various authors have recently probed LLMs for cooperative AI capabilities (Aher et al., 2023; Chan et al., 2023a). The former work examines to what extent LLMs can simulate the behaviour of diverse human subjects in experiments inspired by studies in social psychology, linguistics and behavioural economics. The latter work collects a data set of text descriptions corresponding to archetypal game-theoretic scenarios and evaluates the decisions taken by LLMs in those settings in comparison with human behaviour.

Thus, the stage is set for a rigorous empirical study of the cooperative AI problem. We invite practitioners to put it on the same footing as user-alignment as a pre-eminent focus for the safe, ethical and beneficial deployment of advanced AI assistant technology.

Chapter 15

Access and Opportunity

A. Stevie Bergman, Renee Shelby, Iason Gabriel

Synopsis: With the capabilities described in this paper, advanced AI assistants have the potential to provide important *opportunities* to those who have access to them. At the same time, there is a risk of *inequality* if this technology is not widely available or if it is not designed to be accessible and beneficial for all. In this chapter, we surface various dimensions and situations of *differential access* that could influence the way people interact with advanced AI assistants, case studies that highlight risks to be avoided, and access-related challenges need to be addressed throughout the design, development and deployment process. To help map out possible paths ahead, we conclude with an exploration of the idea of *liberatory access* and look at how this ideal may support the beneficial and equitable development of advanced AI assistants.

15.1. Introduction

With the development of the sophisticated capabilities described in this paper (see Chapter 4), advanced AI assistants hold the potential to greatly improve the opportunities of those who have good access to them, particularly if these technologies support or perform an increasing range of interpersonal and institutional activities. However, we also live in a world shaped by interlocking inequalities (Combahee River Collective, 1977; Crenshaw, 1989; Hill Collins, 2009) where access to opportunities, goods and technologies is often unevenly distributed and shaped by hierarchies including those pertaining to gender, sexuality, disability, religion, race and ethnicity. In a future where the use of advanced AI assistants is a boon for users, but where there has not been intentional design for equitable access (Costanza-Chock, 2020; Davis et al., 2021; Ovalle et al., 2023; Rigot, 2022), we risk encountering a divide between the 'haves' and 'have-nots' that operates across these different dimensions of opportunity and marginalisation. Failure to consider existing social structures, and the ways they interact with the design and function of technology, could result in forms of AI that increase access in some areas while simultaneously compounding asymmetric or harmful relationships in other walks of life (Bennett and Keyes, 2020; Tucker, 2017).

As we show below, the process of designing more general AI assistants – and even assistants that target beneficial use cases in high impact domains, such as health, education or economic empowerment (Chui et al., 2018) – may not be sufficient to ensure that the technology performs well for marginalised communities or to provide these groups with high-quality access to the opportunities it creates. To achieve these outcomes, further steps are needed, including those that speak to the specific needs and experiences of these communities and to the kinds of obstacle they encounter (Mingus, 2017; Zajko, 2022; Bennett et al., 2018; see Chapter 19). The more holistic approach proposed here requires methods that reach beyond 'computational correctives [that] invariably fall short' when viewed as the complete answer to complex societal issues (Davis et al., 2021, 1).

Instead, efforts to support widespread opportunity and access need to attend to various power dynamics that shape how technologies, including AI assistants, are developed and experienced by different communities and users – and to take measures to address these dynamics when envisaging their deployment and use.

To achieve positive outcomes, advanced AI assistants must be designed both *for* and *with* historically marginalised communities to ensure that they are properly responsive to the needs of those who have sometimes been pushed to the periphery (Bennett et al., 2020). Too often, technological pathways involving access to resources and opportunity have been disproportionately designed by and for people who are not fully representative of the societies that these technologies are situated within (Henrich et al., 2010; Linxen et al., 2021). Altering this pattern requires identifying levers or processes for representation and inclusion that can affect meaningful change (Combahee River Collective, 1977; Crenshaw, 1989) and intentionally designing with the needs of everyone in mind (Benjamin, 2020; Rigot, 2022). Without such praxis, power imbalances that operate at a societal level have the potential to be replicated via inequitable access to and through advanced AI assistants. By way of contrast, AI systems that proactively tackle these deeper disparities hold out hope for greater equity in access and in the distribution of benefits. We add nuance to this picture below. For now, we begin by assuming that advanced AI assistants will be presumptively beneficial to those with full access to them ('users').

15.2. Inequality and Technology

Inequitable access to technology refers to disparities in how digital technologies are structured, accessed and used (Kvasny, 2006). As digital technologies are increasingly part of the infrastructure needed for meaningful participation in social, economic, and political life (OECD, 2018a; UN, 2021), inequitable access to technology may lead to social exclusion in these and other domains affecting material well-being (Robinson et al., 2015) (see Chapter 6). The stakes involved in equitable access are high: 'information technology, and the ability to use it and adapt it, is a critical factor in generating and accessing wealth, power and knowledge in our time' (Castells, 2009, 92). Scholars have mapped out the economic, political and social harms of digital inequality, including barriers to economic and educational opportunities and cultural capital (van Dijk, 2006). Moreover, the *distribution* of these harms is often patterned along existing axes of inequality, including gender, race and ethnicity, class, disability and nationality, among others (Bennett et al., 2018; Mesch and Talmud, 2011; Ono and Zavodny, 2008; Wasserman and Richmond-Abbott, 2005; Witte and Mannon, 2010). In this way, the (in)accessibility of technology is constitutively intertwined with power and opportunity at the societal level.[1]

Addressing inequitable access to technology requires researchers, developers and policymakers to move beyond 'technosolutionist' paradigms that focus solely on the availability or ownership of a technological device as the answer to social problems (Warschauer, 2002, 2004). Indeed, earlier work on inequitable access has sometimes been critiqued for its binary focus on a one-dimensional frame of 'haves' and 'have-nots'. This focus tends to ignore the dynamic nature of digital inequality (DiMaggio and Hargittai, 2023; Warschauer and Matuchniak, 2010), which extends to include motivational access (e.g. wanting/trust in technology), physical access (e.g. owning a device), skills (e.g. understanding how to use a technology) and use (e.g. ways of using a technology) (van Dijk, 2006). As a consequence, people may be able to access technology in one sense but not another, thus frustrating its potential for productive use. A more holistic view of 'access' enables a shift away from interventions focused only on device ownership as a pathway for access (e.g. the widely criticised 'One Laptop Per Child' initiative, Keating, 2009),[2] towards recognition that fostering meaningful and collective

[1]The authors understand that 'accessibility' is a term most-often used in the US to refer to accessibility for people with disabilities. However, we employ a more expansive notion of the term here to reflect a wide range of sociopolitical factors shaping 'access to or inaccessibility to engage with technology' (Bjørn et al., 2023, 88).

[2]While the programme initially received much positive attention and was endorsed by the United Nations Development Programme

access requires deeper attention to the ways technology is designed, deployed and experienced with equity and social justice in mind. Ultimately, we believe that technological access is best thought of as a kind of social relationship involving 'bundles and webs of powers that enable actors to gain, control and maintain access' to opportunities and goods (Ribot and Peluso, 2003, 154).

Questions about access to technology fundamentally concern social norms and expectations about embodied ways of being in the world. All technologies are reflections of a world view (Alkhatib, 2021), and the ways they are designed and deployed says something about who they are meant to serve and about who belongs and who does not. As disability studies scholar Titchkosky (2011, 6) notes: 'exploring the meanings of access is, fundamentally, the exploration of the meaning of our lives together – who is together with whom, how, where, when and why'. Through this lens, the relational notion of 'access' offers an entry point for deeper consideration of how technology may be shaped by social norms that foreground the needs, priorities and perspective of some user groups – for example those who are able-bodied (Bennett et al., 2018; Gregor et al., 2005) – over others. Given that questions of 'access' implicate how differently situated communities interact and relate to one another, 'access' can be productively employed as a lens for examining what kinds of relations technologies engender at the societal level.

15.3. Case Studies: Access, Opportunity and AI

As numerous authors have shown, AI technologies are not value neutral, often serving to shape opportunities, pathways and dependencies across the contexts in which they are deployed (Amoore, 2020; Broussard, 2023; Chun and Barnett, 2021). Furthermore, all AI technologies are reflections of the social world, embodying choices made by developers in AI pipelines (Alkhatib, 2021; Ovalle et al., 2023; Suresh and Guttag, 2021). A dominant paradigm for developing technology is to design for an imagined 'biggest use case' scenario, which most often focuses on user communities that are located in white, middle- and upper-class, Western-centric contexts (Rigot, 2022). Much existing research documents the limitations of this development paradigm, including through the lenses of gender (Bivens and Haimson, 2016; Noble, 2020), disability (Bennett and Keyes, 2020; Morris, 2020; Whittaker et al., 2019), race (Benjamin, 2020; Sweeney, 2013) and geopolitical context (Kak, 2020; Png, 2022; Sambasivan et al., 2021). Together, these studies illustrate how a specific vision of product development can embed inequities into AI systems leading to unequal performance and access in the real world (DeVito et al., 2021; Dorn, 2019; Harrington et al., 2022; Martin and Wright, 2023). These insights underscore the need to address inequality in technology development by designing for and with the margins. Indeed, ignoring the impact on different communities and user groups risks falling into a 'false universalism' which benefits some demographics at the expense of others (Roberts and Jesudason, 2013, 315). In the remainder of this section, we discuss two case studies involving potentially unequal access: *AI voice assistants* and *required use* of digital technologies.

Disparate performance of AI voice assistants based on identity

Inferior *quality of access* to technology, for certain user communities, has been uncovered for a wide range of AI technologies that rely on biometric data (e.g. facial features, skin tone or voice) as system inputs, including computer vision (Buolamwini and Gebru, 2018; Raji et al., 2020a) and speech recognition systems (Koenecke et al., 2020; Mengesha et al., 2021). AI technologies with disparate performance potentially lead to quality-of-service harms, which unevenly distribute the social benefits of a technology and may lead to users experiencing alienation, additional labour and service or benefits loss when access is of limited quality (Shelby et al.,

(A.P., 2006), it largely failed due to the sociocultural mismatch between the expectations and views of what laptops can achieve in Western nations versus the context and expectations in which they were deployed in the Global South (McArthur, 2009).

2023). For instance, Mengesha et al.'s (2021) study of AI voice assistant interaction with users who speak African-American Vernacular English found that the failures of these systems for this speaker group have: (1) behavioural ramifications that influence how users adapt to technology (e.g. changing how they speak) and (2) psychological impacts, such as feelings of alienation, when users recognise that a technology is failing them because of their identity.

In these cases, inferior quality of access has the potential to exacerbate existing social inequalities (e.g. Kazenwadel and Steinert, 2023), for example, those regarding race (Bonilla-Silva, 1997; DuBois and Eaton, 1996; Feagin, 2013, 1991; Massey and Denton, 1993). Disparate performance of technologies along lines of race is one mechanism through which systemic racism may be enacted in the digital domain (Zalnieriute and Cutts's report to the UN Human Rights Council, 2022). To accommodate inferior quality of access in practice, a user may 'code switch', meaning they have to use a different language, dialect or accent to make the technology work or improve its performance (Mengesha et al., 2021). In the words of one participant: 'because of my race and location, I tend to speak in a certain way that some voice technology may not comprehend. When I don't speak in my certain dialect, I come to find out that there is a different result in using voice technology' (Mengesha et al., 2021, 7). This situation of differential access leads to lower utility and worse outcomes for these users. Users adapting their speech instead of using their native dialect also makes it harder for developers to detect when the device works poorly for a given language variety without direct consultation or the development of new feedback mechanisms (Birhane et al., 2022; Weidinger et al., 2021).

Required access: The impacts of digital technology as societal infrastructure

When access to a particular digital technology is *required to engage* with an organisation (e.g. governments (Thiel, 2020), social services (Eubanks, 2006) or humanitarian aid (Iazzolino, 2021)), access-related concerns intertwine with issues of consent, autonomy, and surveillance. For instance, information infrastructure enables processing of personal data and protocols to 'mediate between individuals and the organisations with which we relate' (Lyon, 2008, 500). The 'access' lens offers a way to trace and reveal the politics of an information infrastructure. For instance, India's Aadhaar (unique identity number) software requires facial, iris and fingerprint scans to access financial entities (e.g. Amazon Pay) and Indian government benefits (Macdonald, 2023), open a business or register for a goods and services tax number (Burt, 2023). At the US border, migrants who do not use the US Customs and Border Protection app, CBP One, encounter a higher bar for claiming asylum (Gottesdiener et al., 2023). The Atlantic Plaza Towers complex in New York City, a rent-controlled apartment complex with predominantly Black residents, required the use of facial recognition to access the building (Moran, 2020) despite such systems' documented performance inequities across axes of gender and race (Buolamwini and Gebru, 2018). In these examples, the lens of 'access' illustrates how refusal to participate carries financial, legal and social risks. However, even if someone wants to use a particular technology, if the system performs poorly for them, it may carry the same punitive consequences of opting out.

The above case studies illustrate different dynamics that could influence how communities experience advanced AI assistants across a range of contexts. To address these risks, and ensure that existing inequalities are not compounded, design interventions geared towards widespread access and opportunity are needed (Helsper, 2021; see Chapter 14). We discuss the challenges and opportunities created by advanced assistants in more detail below.

15.4. Access and Advanced AI Assistants

Experiences of differential access to technology can take many forms, including total lack of access, inferior quality of access and access to an actively bad, misaligned or punitive system.[3] Indeed, one or more of these situations of differential access may occur simultaneously, depending on the context. To illustrate the various manifestations of unequal access, we describe below a suite of situations which emphasise the social impacts and type of access that other, differently situated communities may have.

Situation type 1: First, we consider differential access situations where the advanced AI assistant is *beneficial to those who have access*, but other people do *not have any access* to the assistant, as they are totally shut out from use (e.g. via paywalls or need for an institutional affiliation to gain access). In such circumstances, the cost of missed opportunities from access to advanced AI assistants could be either consequential or inconsequential to those without access. For example, this could be an education assistant that helps those who have access to write better college application essays and thus obtain better educational opportunities (Singer, 2023). It may also be an assistant that helps those who have it to gain faster and easier access to basic goods and services, such as healthy food and government benefits. Those without access to the assistant must spend more time and resources to obtain the same goods.[4] Alternatively, there are situations where — while only a subset of society may have access — there are still beneficial knock-on effects to those who do not have access (e.g. an advanced AI assistant that spurs socially beneficial scientific or medical research,[5] one that supports democratic institutions by ensuring politicians are held accountable to the citizenry or another that reduces the prevalence of misinformation) (e.g. Harutyunyan, 2023). In such cases, while only some researchers may have access, they could make discoveries that benefit those without access to the technology, thus creating positive spillover effects (see Chapters 14 and 17).[6] Even in these scenarios, existing social inequalities may shape the distribution and flow of potential AI-assisted discoveries, as was seen with inequalities in the distribution of Covid-19 vaccines[7] (Tatar et al., 2022, 2).

Situation type 2: The second type of situation to consider is where some users experience full, beneficial and high-quality access while others have *inferior quality of access* to an otherwise good assistant. This is a classic, inequitable quality-of-service situation. These user experiences may include slower or less advanced access, where the access is inferior due to, for example, slower internet speeds (e.g. net neutrality; see Finley, (2020) or other infrastructural differences leading to the 'digital divide'[8]) or a lower tier/free version of a paid subscription (e.g. Rogers, 2023; Shankland, 2023). In this case, the technology would perhaps take significantly longer to load, join queues and so on (see Chapter 14). The user can still access most services, but time costs are elevated or there is limited availability of some capabilities (e.g. plugins). Opportunities may be missed due to these issues. Alternatively, some users could experience less smooth or buggier access to the assistant. In these situations, the assistant may not be able to execute every necessary task, or it may be less-aligned with the users' intentions (see Chapter 5). As an example of how this could occur, we could envision an advanced AI assistant suffering from a disparate performance issue seen in modern-day digital assistants (e.g. Mengesha et al., 2021; Lima et al., 2019; Wu et al., 2020) where English speakers with a regional or non-American accent are unable to access time-saving services such as to-do lists, etc., as the assistant does not understand the

[3] While in this paper we focus on inequitable access, we could certainly imagine alternative situations to those on the above list, in which those with 'good' access experience some benefits in conjunction with wider destructive effects (e.g. addiction to or overdependence on the technology).

[4] For example, NowPow.

[5] A current day example is, perhaps, the AlphaFold Protein Structure Database (AlphaFold).

[6] However, note that the diversity and experiences of researchers is consequential to the applicability of their research and problem areas they pursue (e.g. Harding, 1986; Longino, 1993).

[7] For example, Pfizer has characterised the development of Covid-19 vaccines as AI-assisted (Pfizer, 2023).

[8] The 'digital divide' in its simplest formulation refers to the 'gap between those with Internet access and those without it' (Muller and Aguiar, 2022).

users' speech. In such situations, the user may have a frustrating experience, but the assistant is generally still usable and useful. These users may experience common quality-of-service harms, as delineated by Shelby et al. (2023).

Situation type 3: In the last set of situations, some users *only have access* to an *actively bad, punitive or misaligned* AI assistant, meaning that the technology is not merely frustrating or inconvenient to use, but has actively negative effects for these users and/or society (e.g. forms of misalignment, or mistakes made by the assistant, have a punitive impact on the user or a resounding societal impact) (see Chapters 5, 7 and 8). An example could be one that directly mirrors present-day issues with facial recognition use in policing in the US (Johnson and Johnson, 2023). An advanced AI assistant could be employed by police officers that is less capable of distinguishing between Black profiles, thus causing the wrong individuals to be apprehended (Johnson, 2022). This situation would in one respect be bad for police officers (the users, in this case) because they end up apprehending the wrong person and presumably generate a negative public reaction or expose municipalities to civil suits (Benedicto, 2023; Thanawala, 2023). However, the situation would be still worse for society and those incorrectly arrested (presumably non-users). As the introduction of algorithmic and big data tools in police departments have already shown, they can reproduce and deepen patterns of social inequality and power imbalances (Brayne, 2021; Ferguson, 2017). This example underscores the need to consider the broader context in which advanced AI assistants will be used.

Notably, the effects of these access issues may be *detectable to the user*,[9] as in the police facial recognition example described above, or they may be *undetectable* if the error happens 'behind the scenes', thus potentially leaving the user or those impacted in a state of uncertainty or confusion. This could potentially occur in job application and hiring situations, where the user is not offered a position but cannot know if that was because they were not the right fit or because they had a faulty/misaligned AI assistant (Bogen, 2019). Alternatively, if the advanced AI assistant is poor at providing factual information in a non-English language, it could create an experience that contains errors or misinformation (see Chapter 16). Indeed, misinformation generated in this way may be subtle and not easily detectable by the user. This case reflects the present-day challenges encountered by content moderation on social media, where journalists' accounts have sometimes been removed by mistake and monitoring systems have proved less capable of detecting non-English misinformation (e.g. Fatafta, 2021; Avaaz, 2020). For an advanced AI assistant with greater capabilities, malignant errors that are hard to detect could be more pronounced or spread at a greater rate, and they could lead to the assistant carrying out tasks that are damaging to the the user or their broader community (see Chapter 7).

In the *differential access* situations described above, the groups of people who do not have full or fully beneficial access may be either *randomly* or *systematically* distributed across society. Random distribution might not be due to any systemic societal issues but to, for example, the inherent probabilistic errors that an advanced AI assistant will make (Bommasani et al., 2022a; Creel and Hellman, 2022). For example, generative AI systems often contain conventional machine-learning (ML) models (e.g. binary or multi-class classifiers employed as input/output filters) for automated content moderation strategies to align to established product policies (Solaiman, 2023). These safety classifiers might take down some accounts at random. Alternatively, safety classifiers may disproportionately fail for certain social groups in a way that is correlated with the manner in which that speaker group uses language (e.g. Dias Oliva et al., 2021), and the errors may be *systematically* distributed, necessitating specific interventions designed to address this effect (e.g. Hao et al., 2023).

In general, more systematically distributed access restrictions occur when lack of – or poorer quality – access correlates with other vectors of exclusion such as race, class, disability, living in the Global South or speaking

[9] A second example could be analogous to a situation described by Eubanks (2017) where, to access social services, welfare claimants are forced to use an assistant rather than speak to a human. The effect of this erodes interpersonal relationships between claimants and caseworkers, and it increases denial of social services through errors in automated decision-making.

a language other than English. Furthermore, the situation types described in this chapter illuminate how in the case of an advanced AI system, inaccessibility can either be *direct*, meaning the inaccessibility is to the advanced AI assistant itself, or *indirect*, where the primary point of inaccessibility is to other goods, services and opportunities. These distinctions are important to recognise as we seek both mitigations to inequities and to facilitate broad access to resources. In particular, a direct access issue calls our attention to the barriers that may be preventing access to the technology itself (e.g. not being able to purchase a smartphone, or other examples of what (van Dijk, 2006) calls good physical access, whereas problems of indirect access often require action in relation to the opportunities and affordances that access to the technology provides.

With any given technology, multiple *situations of differential access* can be at play at once, and they can reproduce social divisions and unequal material outcomes without proactive mitigation. This also holds for advanced AI assistants. We next consider access-related risks that may arise for advanced AI assistants, alongside an understanding of current technical capabilities, before discussing potentially emergent access-related risks for more capable technologies.

15.5. Access-Related Risks and Advanced AI Assistants

The most serious access-related risks posed by advanced AI assistants concern the entrenchment and exacerbation of existing inequalities (World Inequality Database) or the creation of novel, previously unknown, inequities. While advanced AI assistants are novel technology in certain respects, there are reasons to believe that – without direct design interventions – they will continue to be affected by inequities evidenced in present-day AI systems (Bommasani et al., 2022a). Many of the access-related risks we foresee mirror those described in the case studies and types of differential access. In this section, we link them more tightly to elements of the *definition* of an advanced AI assistant to better understand and mitigate potential issues – and lay the path for assistants that support widespread and inclusive opportunity and access. We begin with the existing capabilities set out in the definition (see Chapter 2) before applying foresight to those that are more novel and emergent.

Current capabilities: Artificial agents with natural language interfaces

Artificial agents with *natural language interfaces* are widespread (Browne, 2023) and increasingly integrated into the social fabric and existing information infrastructure, including search engines (Warren, 2023), business messaging apps (Slack, 2023), research tools (ATLAS.ti, 2023) and accessibility apps for blind and low-vision people (Be My Eyes, 2023). There is already evidence of a range of sociotechnical harms that can arise from the use of artificial agents with natural language interfaces when some communities have inferior access to them (Weidinger et al., 2021). As previously described, these harms include inferior quality of access (in situation type 2) across user groups, which may map onto wider societal dynamics involving race (Harrington et al., 2022), disability (Gadiraju et al., 2023) and culture (Jenka, 2023). As developers make it easier to integrate these technologies into other tools, services and decision-making systems (e.g. Marr, 2023; Brockman et al., 2023; Pinsky, 2023), their uptake could make existing performance inequities more pronounced or introduce them to to new and wider publics.

At the same time, and despite this overall trend, AI systems are also not easily accessible to many communities. Such direct inaccessibility occurs for a variety of reasons, including: purposeful non-release (situation type 1; Wiggers and Stringer, 2023), prohibitive paywalls (situation type 2; Rogers, 2023; Shankland, 2023), hardware and compute requirements or bandwidth (situation types 1 and 2; OpenAI, 2023), or language barriers (e.g. they only function well in English (situation type 2; Snyder, 2023), with more serious errors occurring in other languages (situation type 3; Deck, 2023). Similarly, there is some evidence of 'actively bad' artificial agents

gating access to resources and opportunities, affecting material well-being in ways that disproportionately penalise historically marginalised communities (Block, 2022; Bogen, 2019; Eubanks, 2017). Existing direct and indirect access disparities surrounding artificial agents with natural language interfaces could potentially continue – if novel capabilities are layered on top of this base without adequate mitigation (see Chapter 3).

Novel capabilities: Access-related risks for advanced AI assistants

AI assistants currently tend to perform a limited set of isolated tasks: tools that classify or rank content execute a set of predefined rules or provide constrained suggestions, and chatbots are often encoded with guardrails to limit the set of conversation turns they execute (e.g. Warren, 2023; see Chapter 4). However, an artificial agent that can *execute sequences of actions on the user's behalf* – with 'significant autonomy to plan and execute tasks within the relevant domain' (see Chapter 2) – offers a greater range of capabilities and depth of use. This raises several distinct access-related risks, with respect to liability and consent, that may disproportionately affect historically marginalised communities.

To repeat, in cases where an action *can only be executed* with an advanced AI assistant, not having access to the technology (e.g. due to limited internet access, not speaking the 'right' language or facing a paywall) means one cannot access that action (consider today's *eBay* and *Ticketmaster* bots). Communication with many utility or commercial providers currently requires (at least initial) interaction with their artificial agents (Schwerin, 2023; Verma, 2023a). It is not difficult to imagine a future in which a user needs an advanced AI assistant to interface with a more consequential resource, such as their hospital for appointments or their phone company to obtain service. Cases of inequitable performance, where the assistant *systematically performs less* well for certain communities (situation type 2), could impose serious costs on people in these contexts.

Moreover, advanced AI assistants are expected to be designed to act in line with user *expectations*. When acting on the user's behalf, an assistant will need to infer aspects of what the user wants. This process may involve interpretation to decide between various sources of information (e.g. stated preferences and inference based on past feedback or user behaviour) (see Chapter 5). However, cultural differences will also likely affect the system's ability to make an accurate inference. Notably, the greater the cultural divide, say between that of the developers and the data on which the agent was trained and evaluated on, and that of the user, the harder it will be to make reliable inferences about user wants (e.g. Beede et al., 2020; Widner et al., 2023), and greater the likelihood of performance failures or value misalignment (see Chapter 11). This inference gap could make many forms of indirect opportunity inaccessible, and as past history indicates, there is the risk that harms associated with these unknowns may disproportionately fall upon those already marginalised in the design process.

Emergent access risks for advanced AI assistants

Emergent access risks are most likely to arise when current and novel capabilities are combined. Emergent risks can be difficult to foresee fully (Ovadya and Whittlestone, 2019; Prunkl et al., 2021) due to the novelty of the technology (see Chapter 1) and the biases of those who engage in product design or foresight processes D'Ignazio and Klein (2020). Indeed, people who occupy relatively advantaged social, educational and economic positions in society are often poorly equipped to foresee and prevent harm because they are disconnected from lived experiences of those who would be affected. Drawing upon access concerns that surround existing technologies, we anticipate three possible trends:

- **Trend 1: Technology as societal infrastructure.** If advanced AI assistants are adopted by organisations or governments in domains affecting material well-being, 'opting out' may no longer be a real option for

people who want to continue to participate meaningfully in society. Indeed, if this trend holds, there could be serious consequences for communities with no access to AI assistants or who only have access to less capable systems (see also Chapter 14). For example, if advanced AI assistants gate access to information and resources, these resources could become inaccessible for people with limited knowledge of how to use these systems, reflecting the skill-based dimension of digital inequality (van Dijk, 2006). Addressing these questions involves reaching beyond technical and logistical access considerations – and expanding the scope of consideration to enable full engagement and inclusion for differently situated communities.

- **Trend 2: Exacerbating social and economic inequalities.** Technologies are not distinct from but embedded within wider sociopolitical assemblages (Haraway, 1988; Harding, 1998, 2016). If advanced AI assistants are institutionalised and adopted at scale without proper foresight and mitigation measures in place, then they are likely to scale or exacerbate inequalities that already exist within the sociocultural context in which the system is used (Bauer and Lizotte, 2021; Zajko, 2022). If the historical record is anything to go by, the performance inequities evidenced by advanced AI assistants could mirror social hierarchies around gender, race, disability and culture, among others – asymmetries that deserve deeper consideration and need to be significantly addressed (e.g. Buolamwini and Gebru, 2018).
- **Trend 3: Rendering more urgent responsible AI development and deployment practices**, such as those supporting the development of technologies that perform fairly and are accountable to a wide range of parties. As Corbett and Denton (2023, 1629) argue: 'The impacts of achieving [accountability and fairness] in almost any situation immediately improves the conditions of people's lives and better society'. However, many approaches to developing AI systems, including assistants, pay little attention to how context shapes what accountability or fairness means (Sartori and Theodorou, 2022), or how these concepts can be put in service of addressing inequalities related to motivational access (e.g. wanting/trust in technology) or use (e.g. different ways to use a technology) (van Dijk, 2006). Advanced AI assistants are complex technologies that will enable a plurality of data and content flows that necessitate in-depth analysis of social impacts. As many sociotechnical and responsible AI practices were developed for conventional ML technologies, it may be necessary to develop new frameworks, approaches and tactics (see Chapter 19). We explore practices for emancipatory and liberatory access in the following section.

15.6. Beyond Mitigation: From Unequal to Liberatory Access

The (in)accessibility of technology is constitutively intertwined with social inequality. Conversely, meaningful access to technology can be understood as a way of challenging inequality and a way of enabling productive cooperation between individuals understood as equals (Anderson, 1999). As we have sought to elucidate in this chapter, questions of 'access' implicate how differently situated communities interact and relate to one another through technology. We have also shown how 'access' can be employed as a lens through which to examine kinds of social power relations technologies engender. Against this backdrop, disability justice scholar Mingus (2017) writes: 'liberatory access calls upon us to create different values…and demands that the responsibility for access shifts from being an individual responsibility to a collective responsibility'. Liberatory access provides a goal and set of methods for developing sociotechnical systems that embody the kinds of social relations that challenge social inequalities and support mutual flourishing.

One approach to developing technology that embodies liberatory access is by designing for the margins. *Design for the margins* (DFM) is an approach to design that centres the most impacted and marginalised users from ideation to production (Rigot, 2022). Conventional design processes sometimes view marginalised users as 'edge cases' whose needs are framed as different or 'extra'. Their needs are commonly retrofitted to an

already designed technology. By way of contrast, DFM places their needs at the centre of the design process to dismantle systems of interlocking inequality (Hill Collins, 2009). Crucially, users on the periphery who often receive the least support tend to possess deep experiential knowledge of how to improve technologies that can be better for everyone (Rigot, 2022). DFM offers a methodology for rethinking approaches to 'participation' in product development – with the goal of bringing this knowledge back in and explicitly centering those who are marginalised in a given context. Overall, this approach tends to foster more equitable and safe technologies,[10] including identification of necessary interventions such as inclusive education and customisable interfaces, so that the technology can be fully responsive to marginalised users' needs.

Placing these communities at the centre of the design process opens possibilities for more inclusive, liberatory technologies. Many participatory design approaches have been critiqued as *extractive* insofar as they engage communities through consultation without attention to context (Sloane et al., 2022) or through 'tokenistic forms of 'voice' that fail to redistribute power and agency' (Ymous et al., 2020). The aspiration to embed participation-as-justice proceeds differently and involves developing long-term relationships 'based on mutual benefit, reciprocity, equity and justice' (quoted in Suresh et al., 2022, 667). Importantly, DFM centres those most impacted in the design process while being attentive the broader sociopolitical and institutional contexts in which technologies operate and exist. This requires rethinking the design process in potentially significant ways. As described by Rigot (2022), implementing this approach requires: (1) identifying and prioritising communities who bear the most risk and have the least protection, and bringing in facilitators who can identify those communities and have trusted relationships with them, especially when directly involving a particular community in the design process poses safety risks (e.g. Bellini et al., 2023); (2) zooming in on social, legal and political issues that arise within that particular context; (3) centring the needs of communities from ideation through development, not just including them at later stages of a product development life cycle after numerous decisions have already been made; and (4) regeneralising so that these findings can be scaled and applied alongside insights from user groups who typically are centred in product development (Rigot, 2022, 65). By centring the margins, DFM offers the kind of *transformative approach to technology* and the design of sociotechnical systems that can facilitate liberatory access in service to the goals of opportunity, access and liberation (Mingus, 2017).

15.7. Conclusion

Questions about access to technology fundamentally concern social norms and expectations about how communities interact and relate to one another. This chapter has discussed how 'access' can be employed as a lens for examining the kinds of relations technology engenders in ways that extend beyond questions of merely technical and logistical access. Both the developers of advanced AI assistants *and* the organisations that adopt them have a responsibility to consider these relationships and to assess and mitigate access-related risks. For developers, attention must be given to how assistants are constructed and who they are optimised for. This requires an understanding of existing access inequalities and concerted engagement with different publics to understand how to address them effectively (Björgvinsson et al., 2010; Dantec and DiSalvo, 2013; Erete et al., 2023). For organisations that adopt advanced AI assistants, attention must be paid to technological limitations and risks of adoption, in addition to their benefits. Moreover, these organisations have a duty to understand how performance limitations of technological agents will interact with the existing access inequalities in their domain that are experienced by their clients or constituents. One way for developers and organisational adopters of advanced AI assistants to approach this work is by employing 'access' as a lens for anticipating

[10] For example, development of a dating app for LGBTQ people might focus on the needs of potential users whose sexuality is criminalised (Article 19, 2018), or considerations of privacy or safety might centre the needs of journalists and activists in parts of the world that are subject to harassment, surveillance, arrest or assassination (Article 19, 2022; Warford et al., 2022).

potential situations of differential access and who might experience them, and by drawing on multidisciplinary best practices, particularly from fields focused on equity and access, such as disability justice (e.g. Berne et al., 2018; Mingus, 2011). These fields can provide the political container necessary for grounding analyses and materially moving towards liberatory access.

Chapter 16

Misinformation

Nahema Marchal, Iason Gabriel, Arianna Manzini, Geoff Keeling, Beth Goldberg, Josh Goldstein

Synopsis: Advanced AI assistants pose four main risks for the information ecosystem. First, AI assistants may make users more *susceptible* to misinformation, as people develop trust relationships with these systems and uncritically turn to them as reliable sources of information. Second, AI assistants may provide ideologically *biased* or otherwise *partial information* to users in attempting to align to user expectations. In doing so, AI assistants may reinforce specific ideologies and biases and compromise healthy political debate. Third, AI assistants may *erode* societal *trust* in shared knowledge by contributing to the dissemination of large volumes of plausible-sounding but low-quality information. Finally, AI assistants may facilitate *hypertargeted disinformation* campaigns by offering novel, covert ways for propagandists to manipulate public opinion. This chapter articulates these risks and discusses technical and policy mitigations.

16.1. Introduction

Recent advances in the field of AI have enabled AI systems to develop unprecedented capabilities, such the ability to generate human-like text, images, video and audio (Nightingale and Farid, 2022; Spitale et al., 2023), to teach themselves how to reason, use external tools and to take actions in the real world (Mialon et al., 2023; Schick et al., 2023).

With these advances come growing concerns about the potential for AI systems to spread misinformation and fuel online influence operations. A recent survey found that three in four Americans are concerned about AI driving mis- and disinformation (Ipsos, 2023), and leading AI labs have also flagged this as a major risk in relation to large language models (LLMs), stating that with wider adoption, advanced AI systems could 'reinforce entire ideologies, worldviews, truths and untruths [...] cement them or lock them in, foreclosing future contestation, reflection and improvement' (OpenAI, 2023d, 9).

As these systems are rapidly integrated into a range of user-facing applications, including virtual AI assistants (Knight, 2023), it is therefore important to think about the unique challenges this particular form factor might pose for the *integrity of our information environment*. How might widespread adoption of highly capable and adaptive AI assistants impact the spread of misinformation and political propaganda? What impact might these technologies have on information retrieval, knowledge and beliefs? How might this impact public discourse, and what can be done to mitigate these threats?

This chapter focuses on the risks posed by AI assistants, both existing and future, with advanced capabilities such as independent reasoning and planning skills. Throughout the chapter, we refer to 'misinformation' and 'falsehoods' as false information, and to 'disinformation' as the spread of false or misleading information with

the explicit intention of causing harm through, for example, coordinated influence operations (Wardle and Derakhshan, 2017).

The rest of the chapter proceeds as follows. Drawing on the communication and psychology literature, Section 16.2 provides an overview of the mechanisms underlying the spread of mis- and disinformation in the digital era. In Section 16.3, we analyse the role of AI systems in enabling these phenomena. Section 16.4 explores the unique risks and challenges posed by advanced AI assistants for our information environment and explores some potential mitigation strategies.

16.2. The Challenge of Misinformation and Disinformation

Misinformation, disinformation and strategic attempts to manipulate public opinion are far from new (Burkhardt, 2017). Information sharing is central to human culture, and rumours and stories that evoke strong emotions tend to spread quickly and gain credibility through social transmission, regardless of their accuracy (Berger, 2011; Berinsky, 2017; Cotter, 2008). Misinformation can originate from many different sources, including individuals, governments and politicians, and history is replete with examples of people strategically deploying lies and falsehoods to advance their own interests and gain political power. As early as ancient Rome, Octavian waged one of the first known disinformation campaigns against Julius Caesar's general Mark Anthony, using 'short, sharp slogans written upon coins' to smear his reputation and win support for his own claim to power (Kaminska, 2017).

However, rapid advances in digital and communication technologies have made it cheaper and easier than ever to create and disseminate false or misleading information at scale. In the past, political information was relayed primarily through traditional media outlets, such as newspapers, television and radio, with editorial oversight mechanisms in place to ensure the quality and accuracy of the information they communicated. By lowering the cost of information production, the advent of the internet and social media has challenged the standing of these gatekeepers (Jungherr et al., 2020). Today, any internet user – within the bounds of their local access, speech and content regulation laws – can generate and distribute their own content over digital platforms, with little editorial oversight. This has led to a proliferation of news sources, many of which lack or intentionally forgo the quality assurance practices of traditional outlets, including fact-checking (Zhuravskaya et al., 2020).

Digitalisation has also accelerated the *speed* and *scale* at which information travels between media, citizens, political actors and the distribution channels at their disposal. Large social media platforms like *X* (formerly *Twitter*), *Facebook* and *TikTok* have ushered in new forms of social interactions which allow people to connect, interact and share user-generated content with others on a global scale. In addition, the algorithms powering these platforms are designed to reward and prioritise content that stimulates engagement from others (Bakshy et al., 2015). This prioritisation, coupled with the dynamics of networked communication, means that false, misleading or emotion-laden content published on social media has the potential to quickly cascade and reach millions of users quasi-instantaneously (Vosoughi et al., 2018).[1]

Taken together, these developments have created new opportunities for malicious actors to misuse digital tools for political and economic gains (see Chapter 8). Fabricating and spreading false news stories on social media has become a lucrative business in many parts of the world (Hughes and Waismel-Manor, 2021). Online influence operations, which are coordinated attempts by state and non-state actors to influence domestic or foreign politics, have also gained ground across the world over the past decade (Bradshaw and Howard, 2019),

[1] In 2020, for example, a conspiracy video about the origins of the coronavirus pandemic, called 'Plandemic' went viral on social media, racking up eight million views across platforms within days of its release (Frenkel et al., 2020).

spawning a sprawling misinformation-for-hire industry. To achieve their aims, propagandists employ a range of tactics and techniques that often exploit the affordances and vulnerabilities of the online information ecosystem (The Cybersecurity and Infrastructure Security Agency, 2022). These include, for example, exploiting data voids[2] to push people towards false news sites and misleading content, cultivating inauthentic online personas to lend credibility to their narratives (e.g. fake 'experts'), deploying bots – fake profiles that appear to be real individuals – to amplify or drown out political messages in a coordinated manner or fake grassroots support for a specific cause, a technique known as 'digital astroturfing' (Gorwa and Guilbeault, 2020; Woolley, 2016).[3]

What are the consequences of this? Mis- and disinformation pose a number of threats to democracy. Well-functioning democracies require a well-informed citizenry that is able to make informed political decisions on public issues (MacKenzie and Bhatt, 2020). A society in which many people are misinformed or hold beliefs that go against established facts is therefore concerning, not only because it can have negative impacts on individuals but also because it may have harmful repercussions for society as a whole, such as stoking divisions and eroding trust in established truths. There is compelling evidence, for example, that misinformation was instrumental in fuelling violent insurrections in India,[4] stoking racial hatred in Myanmar and discrediting public health sources during the Covid-19 pandemic, leading to millions of preventable deaths (Burki, 2019; Mozur, 2018; Rocha et al., 2023). Multiple studies also link exposure to fake news and unreliable websites with reduced trust and negative attitudes towards mainstream news sources, heightened partisan animosity and decreased interpersonal trust (see Guess et al., 2020a; Hameleers et al., 2022; Ognyanova et al., 2020).

Nevertheless, the prevalence of misinformation on social media remains a contentious issue, with some scholars claiming that its pervasiveness has been exaggerated (see, for example, Allen et al., 2020; Altay et al., 2023; Guess et al., 2020b). The existing literature has yet to determine conclusively whether and how online misinformation shapes political beliefs, and whether these effects are meaningfully different from traditional forms of media influence. Indeed, what makes individuals more susceptible or resistant to misinformation depends on a number of cognitive, social and affective factors such as whether that information confirms or contradicts their pre-existing beliefs and attitudes ('confirmation bias'), their familiarity and trust in the source of information, and the frequency and modality in which they encounter that information ('repetition' and 'elaboration' effects) (for a review, see Ecker et al., 2022). All of these factors are important to take into consideration when considering the challenges that AI assistants might pose for mis- and disinformation.

16.3. Misinformation, Disinformation and AI

Technical advances in AI and machine learning (ML), such as generative AI models and personalised recommender systems, have also created opportunities for boosting the spread of misinformation and disinformation, and raised concerns about these systems' impact on the information ecosystem. First, AI systems have made it easier to generate highly realistic synthetic content that is indistinguishable from real content. This could accelerate the spread of misinformation and prevent truth discernment. Second, AI-powered recommendation systems on digital platforms have enabled more personalised forms of targeting and distribution pathways for misleading content. We will now explore each of these elements in more detail.

[2]Data voids are situations where there is a complete absence of data or lack of reliable or balanced information about a specific topic or query online. These often manifest during breaking news and are easily exploited by malicious actors to promote fringe or conspiratorial content in searches.

[3]For a more comprehensive framework and overview of disinformation tactics, see Pamment (2023) and The Cybersecurity and Infrastructure Security Agency (2022).

[4]In India, for example, false rumours circulated on *WhatsApp* about child kidnappers were implicated in the mob lynching of 29 innocent people (Dixit and Mac, 2018).

Content creation and manipulation

Recent advances in AI image and text generation have expanded the opportunities to produce *natural-sounding text* and *highly realistic synthetic images, videos and audio* (known as 'deep fakes'). While media manipulation was already possible, the advent of widely accessible generative AI tools has made it easier than ever to create and disseminate synthetic content. Research shows that this content is also increasingly undetectable and is easily mistaken as genuine (Nightingale and Farid, 2022; Spitale et al., 2023). These tools have also simplified the manipulation of text, videos and images, thus allowing users to create tailored images, audio or videos for specific purposes. As a result, generative AI systems are already increasingly used in the political sphere to produce deceptive videos and entire 'fake news' websites (Hanley and Durumeric, 2023).

The widespread adoption of generative AI models may pose significant challenges to the integrity of the information ecosystem. Research shows that audiovisual content is more evocative and seen as more persuasive than text (Hameleers et al., 2020; Sundar et al., 2021). Should hyperrealistic and persuasive audiovisual content – such as political deepfakes – become ubiquitous, people may soon be completely unable to discern real from synthetic outputs and therefore be more prone to be misinformed by them. A proliferation of AI-generated content could also have the opposite effect and generate more suspicion and distrust. For example, exposure to deepfakes has been shown to provoke feelings of uncertainty and dampen trust in news (Vaccari and Chadwick, 2020). Lastly, an explosion of synthetic content might increase the mental load on everyday users when navigating online spaces. This could in turn lead to less carefulness, with less concern and other emotional or cognitive effects that are known to boost misinformation sharing and reinforce generalised distrust (Apuke et al., 2022).

As model capabilities expand, there is also a growing concern that these tools could power increasingly sophisticated and harder-to-detect misinformation campaigns to enable new influence tactics (see, for example, Goldstein et al., 2023). Evidence shows that multimodal generative AI tools have already enabled new forms of user manipulation. One such tactic, for example, is 'impersonation' or the ability to speak deceptively on behalf of others by convincingly impersonating them. Scammers are increasingly using AI-generated audio clips to trick people into giving away money and other private or sensitive information (Flitter and Cowley, 2023). Even without the need to deceive audiences directly, an increasing prevalence of synthetic material in the information ecosystem could exacerbate the 'liar's dividend' or the ability of people to dismiss any evidence held against them as fake or AI-generated (Chesney and Citron, 2018). We have already seen examples of this playing out in court (Bond, 2023).

Targeted personalisation

Over the past decade, the rise of AI-powered recommender systems has also transformed how digital information, including misinformation, is distributed and promoted. Today, most digital platforms use algorithmic systems that collect data and personalise the content users see based on their past behaviour, preferences and search results. These systems make recommendations to users to optimise their engagement and activity on the platform (e.g. what video to watch next). Various AI and ML models and techniques, including neural networks, generative adversarial networks (GANs) and reinforcement learning methods have been applied to these systems to augment and personalise user experiences (Zhang et al., 2021).

However, recommender systems have also been criticised for their role in spreading misinformation and conspiracy theories, and for exposing users to increasingly ideologically biased, fringe and radical content, at the risk of leading them to develop extreme views (see Hao, 2021; Tameez, 2020; Tufekci, 2018). Scholars argue that the design of recommender systems can play a significant role in shaping user exposure to this type of content, leading them down a 'rabbit hole' from a more benign to increasingly harmful types of content, such as

content about self-harm and eating disorders (e.g. Ribeiro et al., 2021; Whittaker et al., 2021). Recommender systems also pose clear ethical challenges with respect to user autonomy and inappropriately manipulating or exposing users to risks (Milano et al., 2020). Despite these concerns, recent studies have found that, while recommendation algorithms are influential in shaping users' informational diet, they do not have a measurable impact on their political beliefs (Isaac and Frenkel, 2023).

Another related concern is that malicious actors could weaponise AI tools to supercharge the distribution of microtargeted political propaganda. Online influence operations have become increasingly personalised over the past decade. In the lead-up to the 2016 US election, for example, Russian Internet Research Agency operators used microtargeted ads to cultivate secessionist and nativist sentiments across curated *Facebook* groups and *Instagram* accounts, and to discourage African-American users from voting or supporting specific candidates (DiResta et al., 2019). Politicians and propagandists can now target specific segments of the population based on algorithmically determined filters, including geographical location, consumer preferences, dispositions and behavioural traits (Dommett, 2019). While personalised messaging is not a new concept, AI-powered tools could make it even easier and faster for malicious actors to test different methods and to optimise the timing and tone of their messages for specific individuals, at scale. This type of user profiling poses privacy and anonymity risks, and it raises concerns about voter manipulation[5] (see Chapter 9) and distortion of public discourse (see Bayer, 2020; Milano et al., 2021).

16.4. Misinformation, Disinformation and Advanced AI Assistants

The rapid integration of AI systems with advanced capabilities, such as greater autonomy, content generation, memorisation and planning skills (see Chapter 4) into personalised assistants also raises new and more specific challenges related to misinformation, disinformation and the broader integrity of our information environment. We consider four of them in the section below.

New vulnerabilities to misinformation

First, the specific form factors of assistive AI may increase people's vulnerability to misinformation. Several cognitive, social and political factors influence how people process and respond to information, and their willingness to believe or reject certain claims. Research shows, for example, that people more readily believe information from sources they trust, such as family members and friends (American Press Institute, 2017; Anspach, 2017), and AI assistants can be designed to foster a similar sense of user trust. People tend to perceive autonomous AI systems as competent and trust in their abilities to perform certain tasks (McKee et al., 2021; see Chapter 12). As AI assistants become more personalised and ubiquitous, becoming virtual friends and even romantic partners (Chow, 2023; see Chapter 11), users may develop a high level of trust in them and take the information they provide at face value, even when it is false (Burtell and Woodside, 2023; see Chapters 9, 10 and 12).

Low literacy levels around how content generated by AI assistants is produced and distributed could also complicate user's ability to critically evaluate information. People's ability to discern truth from falsehood varies greatly depending on their level of digital literacy – their understanding and familiarity with how the internet and digital technology work (Sirlin et al., 2021). LLMs – the models currently powering the latest generation of virtual assistants – are known to produce factual inaccuracies and to 'hallucinate', making up semantically plausible but factually inaccurate statements (Ji et al., 2023; Lee et al., 2019). As people, including journalists, adopt AI assistants to assist them with copy-writing, website design or any other tasks involving

[5] Though there is limited evidence of the effectiveness of online political ads on voting behaviour (see Haenschen, 2023, for example).

content generation, the internet might become replete with AI-generated outputs of disputable quality and facticity. Without a clear heuristic understanding of the capabilities and limitations of AI assistants, such as how it summarises information, users may not be equipped to critically evaluate misinformation when they see it.

Entrenched viewpoints and reduced political efficacy

Design choices such as greater personalisation of AI assistants and efforts to align them with human preferences could also reinforce people's pre-existing biases and entrench specific ideologies. Increasingly agentic AI assistants trained using techniques such as reinforcement learning from human feedback (RLHF) and with the ability to access and analyse users' behavioural data, for example, may learn to tailor their responses to users' preferences and feedback. In doing so, these systems could end up producing partial or ideologically biased statements in an attempt to conform to user expectations, desires or preferences for a particular worldview (Carroll et al., 2022). Over time, this could lead AI assistants to inadvertently reinforce people's tendency to interpret information in a way that supports their own prior beliefs ('confirmation bias'), thus making them more entrenched in their own views and more resistant to factual corrections (Lewandowsky et al., 2012). At the societal level, this could also exacerbate the problem of epistemic fragmentation – a breakdown of shared knowledge, where individuals have conflicting understandings of reality and do not share or engage with each other's beliefs – and further entrench specific ideologies.

Excessive trust and overreliance on hyperpersonalised AI assistants could become especially problematic if people ended up deferring entirely to these systems to perform tasks in domains they do not have expertise in or to take consequential decisions on their behalf (see Chapter 12). For example, people may entrust an advanced AI assistant that is familiar with their political views and personal preferences to help them find trusted election information, guide them through their political choices or even vote on their behalf, even if doing so might go against their own or society's best interests. In the more extreme cases, these developments may hamper the normal functioning of democracies, by decreasing people's civic competency and reducing their willingness and ability to engage in productive political debate and to participate in public life (Sullivan and Transue, 1999).

Degraded and homogenised information environments

Beyond this, the widespread adoption of advanced AI assistants for content generation could have a number of negative consequences for our shared information ecosystem. One concern is that it could result in a degradation of the quality of the information available online. Researchers have already observed an uptick in the amount of audiovisual misinformation, elaborate scams and fake websites created using generative AI tools (Hanley and Durumeric, 2023).[6] As more and more people turn to AI assistants to autonomously create and disseminate information to public audiences at scale, it may become increasingly difficult to parse and verify reliable information. This could further threaten and complicate the status of journalists, subject-matter experts and public information sources. Over time, a proliferation of spam, misleading or low-quality synthetic content in online spaces could also erode *the digital knowledge commons* – the shared knowledge resources accessible to everyone on the web, such as publicly accessible data repositories (Huang and Siddarth, 2023). At its extreme, such degradation could also end up skewing people's view of reality and scientific consensus, make them more doubtful of the credibility of all information they encounter and shape public discourse in unproductive ways. Moreover, in an online environment saturated with AI-generated content, more and more

[6] Compounding the issue, as the context surrounding AI-generated content is lost, it may become more difficult to fact-check this content in the future.

people may become reliant on personalised, highly capable AI assistants for their informational needs. This also runs the risk of homogenising the type of information and ideas people encounter online (Epstein et al., 2023).

Weaponised misinformation agents

Finally, AI assistants themselves could become *weaponised* by malicious actors to sow misinformation and manipulate public opinion at scale. Studies show that spreaders of disinformation tend to privilege quantity over quality of messaging, flooding online spaces repeatedly with misleading content to sow 'seeds of doubt' (Hassoun et al., 2023). Research on the 'continued influence effect' also shows that repeatedly being exposed to false information is more likely to influence someone's thoughts than a single exposure. Studies show, for example, that repeated exposure to false information makes people more likely to believe it by increasing perceived social consensus, and it makes people more resistant to changing their minds even after being given a correction (for a review of these effects, see Lewandowsky et al., 2012; Ecker et al., 2022). By leveraging the frequent and personalised nature of repeated interactions with an AI assistant, malicious actors could therefore gradually nudge voters towards a particular viewpoint or sets of beliefs over time (see Chapters 8 and 9).

Propagandists could also use AI assistants to make their disinformation campaigns more personalised and effective. There is growing evidence that AI-generated outputs are as persuasive as human arguments and have the potential to change people's minds on hot-button issues (Bai et al., 2023; Myers, 2023). Recent research by the Center for Countering Digital Hate showed that LLMs could be successfully prompted to generate 'persuasive misinformation' in 78 out of 100 test cases, including content denying climate change (see Chapters 9 and 18).[7]

If compromised by malicious actors, in the future, highly capable and autonomous AI assistants could therefore be programmed to run astroturfing campaigns autonomously,[8] tailor misinformation content to users in a hyperprecise way, by preying on their emotions and vulnerabilities, or to accelerate lobbying activities (Kreps and Kriner, 2023). As a result, people may be misled into believing that content produced by weaponised AI assistants came from genuine or authoritative sources. Covert influence operations of this kind may also be harder to detect than traditional disinformation campaigns, as virtual assistants primarily interact with users on a one-to-one basis and continuously generate new content (Goldstein et al., 2023).

16.5. Risks and Mitigations

As we have discussed, powerful AI assistants have the potential to shape the information environment in significant ways. We summarise below the informational risks these technologies pose at the individual and societal level, and outline a number of possible mitigation strategies that various stakeholders could adopt to reduce those risks.

- **Risk 1: Advanced AI assistants may increase people's vulnerability to misinformation.** First, AI assistants may make users more susceptible to misinformation, as people develop competence trust in these systems' abilities and uncritically turn to them as reliable sources of information.

[7] The researchers also found that LLM-based systems could easily create content in the style of Facebook and X (formerly Twitter) posts, further illustrating their potential for misuse in misinformation campaigns (Center for Countering Digital Hate, 2023; see also Brewster et al., 2023).

[8] The practice of faking grassroots public support for a cause to influence public opinion. In 2017, for example, the industry group Broadband for America generated *millions of comments* with fake or stolen personal information against the Federal Communications Commission's (FCC) proposed reversal of net neutrality (Singel, 2018).

- **Risk 2: Advanced AI assistants may entrench specific ideologies and impact citizens' understanding and engagement with public affairs.** Second, AI assistants may provide ideologically biased or otherwise partial information in attempting to align to user expectations. In doing so, AI assistants may reinforce people's pre-existing biases and compromise productive political debate.
- **Risk 3: Advanced AI assistants erode trust and undermine shared knowledge by polluting the information ecosystem.** Third, AI assistants may contribute to the spread of large quantities of factually inaccurate and misleading content, with negative consequences for societal trust in information sources and institutions, as individuals increasingly struggle to discern truth from falsehood.
- **Risk 4: Advanced AI assistants drive opinion manipulation by empowering online influence operations.** AI assistants may facilitate large-scale disinformation campaigns by offering novel, covert ways for propagandists to manipulate public opinion. This could undermine the democratic process by distorting public opinion and, in the worst case increasing skepticism and political violence.

Several mitigation strategies could help to address those risks. These can be broadly split between technical and policy solutions.

Technical solutions

- **Limit AI assistants' functionality.** To mitigate the risks outlined above, AI developers could introduce limits on AI assistants' functionality at the model level (e.g. prevent them from expressing political opinions). This could be achieved by applying content filters on model outputs or user prompts, or by limiting AI assistants' abilities to learn from external inputs to prevent them from being injected with harmful content from adversarial actors, for example.[9] While effective in the short term, this approach is not foolproof, as powerful AI models can develop emergent capabilities over time and pursue goals which may not have been specified during training (Wei et al., 2022). Addressing this would therefore require continuous monitoring of AI assistants' behaviour and careful evaluation of models' capabilities over time.
- **Develop robust detection mechanisms.** To limit erosion of trust in public knowledge sources, implementing robust mechanisms to detect 'deepfakes' and other misleading media created or propagated by AI assistants will also be crucial. A number of leading AI developers have recently unveiled ML classifiers and watermarking techniques to help with the identification and differentiation of AI-generated text and images from human-created ones (see Abadi et al., 2016; DeepMind, 2023; Intel, 2022). The number of companies providing similar services is growing (see Hsu and Myers, 2023). While promising, these tools are not entirely immune to adversarial attacks and their practical applications remain limited. For example, current text detection tools still perform poorly when dealing with short texts and languages other than English (Sadasivan et al., 2023).[10] Moreover, current methods are not well equipped to deal with real-world cases of misinformation, which are increasingly multimodal (see Hangloo and Arora, 2022). Lastly, open-source models also present a challenge to synthetic media detection, as there tends to be limited control over how these are deployed and used, making it difficult to enforce industry-wide standards and best practices (Theben et al., 2021).
- **Limit personalisation and promote critical thinking.** Another potential solution would be for developers to limit personalisation and embed prompts encouraging critical thinking in the design of AI assistants to ensure users do not get a biased perspective on public issues. Several news organisations and online

[9] Cohere, for example, has undertaken such an approach to harm prevention for its LLMs.

[10] On a test set of English-text, Open AI's classifier only correctly identifies 26% of AI-written text (true positives) as 'likely AI-written', while incorrectly labelling human-written text as AI-written 9% of the time (false positives).

platforms[11] have already experimented with techniques to help users pause and reflect on the content they are exposed to and to encourage them to diversify their information diets (e.g. *Ground News*). Research suggests that such context-aware recommendations and embedded prompts, such as information literacy videos, are effective in inoculating users against misinformation, improve their ability to better identify emotional manipulation and improve their sharing behaviour (Mattis et al., 2022; Roozenbeek et al., 2022). For this strategy to be useful in the case of AI assistants, however, clear transparency would be key to avoiding users' personal autonomy being undermined. Another promising avenue for developers would be to limit encoding political biases into the models powering AI assistants through fine-tuning, by running dedicated evaluations on ideological diversity and diversifying the pool of human raters they solicit feedback from.

- **Emphasise factuality.** It will be imperative for developers to emphasise factuality in AI assistants. This could be achieved by requiring assistants to systematically cite their sources when presenting or retrieving factual claims,[12] for example, and augment their ability to evaluate sources to determine their trustworthiness (Guo et al., 2022b). The task of fact-checkers will become increasingly challenging as they need to sift through vast amounts of AI-generated content, identifying materials that warrant verification and issuing timely corrections. Several fact-checking organisations are already building applications on top of and developing their own LLMs to automate parts of this process. However, it will be important for the entire AI community to invest more capabilities and resources into building fact-checking mechanisms for verifying the accuracy of information presented to users. One pressing challenge for developers in that respect will be to address sociotechnical vulnerabilities both within and outside their systems, including the issue of 'data voids'.[13] To tackle this, models should be trained on diverse and regularly updated data sets, while assistants should be designed to identify these blind spots and inform users about the limitations of available data on certain topics.

Policy solutions

- **Restrict specific uses and applications.** Implementing robust governance mechanisms will also be crucial for mitigating the potential negative impacts of AI assistants on public life. One approach AI companies and policymakers could take is to impose restrictions on explicitly political or malicious uses of AI assistants through enforceable licences and terms of services. These measures could limit the deployment of AI agents for political campaigning, consulting, lobbying or persuading voters to support specific causes or candidates. Recognising similar concerns around misinformation, some social media platforms have previously banned political advertising on their sites. However, scholars have criticised such measures as inadvertently favouring political incumbents who already have ample resources while putting civil society organisations and political newcomers at a disadvantage (Kreiss and Barrett, 2020). Several AI research labs have also taken steps in that direction by explicitly banning the use of their models to disseminate falsehoods, manipulate users or influence politics. However, effective enforcement of these measures relies heavily on the ability to clearly attribute responsibility. As AI assistants become more powerful, acquiring specific objectives or significant autonomy to determine their own actions, how to better identify clear violations of terms of services and take appropriate sanctions against users will be challenging, thus offering promising avenues for future research and governance.

[11] In 2020, Facebook changed how it surfaces content in users' News Feeds and restricted personalised recommendations in an effort to address polarisation (Rosen, 2020).

[12] Although in the future, simply requiring AI models to list their sources may not be sufficient for determining their accuracy, as an advanced model may selectively choose sources that it believes humans will find persuasive.

[13] If a user asks a question in a data void, AI assistants might provide inaccurate or unreliable information relying on alternative sources, as there is simply not enough data to draw from.

- **Implement transparency mechanisms.** Governments and AI companies could implement protocols and policies that require transparency around the use of AI assistants. This could include mandating clear labels for AI-generated content created by assistants, requiring disclosure of, or altogether banning, the use of AI assistants or bots to impersonate or replicate human activity online.[14] AI labs should also be transparent about the demographic composition of the annotators they recruit to fine-tune their models with and about how they evaluate their models for ideological biases. Another idea in this vein would be for AI labs to engage in transparent reporting of harms caused by AI assistants (similar to the AI Incident Database) so that people could track and weigh the benefits and risks of these systems.
- **Support public education.** AI developers and policymakers could support education programmes to raise public awareness about the workings, limitations and biases of AI assistants, and teach people to become more discerning users. As mentioned above, the focus of public education campaigns should be on ways of improving critical thinking skills rather than simply making people broadly aware that AI can be used for disinformation campaigns. Beyond this, programmes could involve, for example, providing training on how to prompt AI assistants efficiently to generate evidence-based high-quality responses or how to fact-check information provided by AI assistants. While there is compelling evidence of the efficacy of digital literacy programmes in certain circumstances, more research is needed to test how effective interventions such as active pre-bunking are in addressing misinformation relayed by personalised AI assistants, for instance.

16.6. Conclusion

Advanced AI assistants pose four main risks for the information ecosystem. First, AI assistants may make users more *susceptible* to misinformation if people develop trusting relationships with these systems and uncritically turn to them as reliable sources of information (see Chapters 9 and 11). Second, AI assistants may provide ideologically *biased* or otherwise *partial information* to users in an effort to align to user expectations. In doing so, AI assistants may reinforce specific ideologies and biases, which in turn will compromise healthy political debate. Third, AI assistants may *erode* societal *trust* in shared knowledge by contributing to the dissemination of large volumes of plausible-sounding but low-quality information. Finally, AI assistants may facilitate *hypertargeted disinformation* campaigns by offering novel, covert ways for propagandists to manipulate public opinion. This chapter articulates these risks and discusses technical and policy mitigations.

[14] California's 2019 legislature passed a law to this effect, the Bolstering Online Transparency (BOT) Act, though the law has been criticised for lacking enforcement mechanisms to incentivise compliance.

Chapter 17

Economic Impact

Conor Griffin, Juan Mateos-Garcia, Sébastien Krier, Geoff Keeling, Alexander Reese, Iason Gabriel

Synopsis: This chapter analyses the potential economic impacts of advanced AI assistants. We start with an analysis of the economic impacts of AI in general, focusing on *employment*, *job quality*, *productivity growth* and *inequality*. We then examine the potential economic impacts of advanced AI assistants for each of these four variables, and we supplement the analysis with a discussion of two case studies: *educational assistants* and *programming assistants*. We conclude with a series of recommendations for policymakers around the appropriate techniques for monitoring the economic impact of advanced AI assistants, and we propose plausible approaches to shaping the type of AI assistants that are deployed and their impact on the economy.

17.1. Introduction

This chapter explores the potential economic impact of advanced AI assistants. We focus on *employment*, *job quality*, *productivity growth* and *inequality*. We first examine how AI in general has impacted these factors, then we explore the potential impacts of advanced AI assistants through two case studies: *educational assistants* and *programming assistants* (see Chapter 4). We conclude by discussing what public policy instruments are available to direct the economic impact of AI assistants towards socially beneficial outcomes.

17.2. How Has AI Affected the Economy to Date?

We can explore the impact of AI on the economy through the lenses of employment, job quality, productivity growth and inequality. These four variables are all important contributors, both positively and negatively, to individual and societal well-being. We analyse the four variables independently, but we note that they are interdependent. For example, the level of labour demand in an economy shapes employee bargaining power and wages, which in turn affects job quality. Even so, each of these factors provides an informative lens through which we can analyse the impact of AI systems on the economy to date. For reasons of scope, we do not analyse the potential effects of AI on other important economic variables such as competition.

Employment

Employment can contribute to well-being positively, for example by providing income (Tay et al., 2017) and a sense of purpose (Bryce, 2018), or negatively.[1] Unemployment, particularly when it is over the long term, is

[1]There are many nuances that may determine, at the individual level, the relative importance of income to well-being, such as the context in which the income was earned and which aspect of well-being is being considered (Pouwels et al., 2008).

associated with a range of ills, including greater risk of committing and suffering from crime (Phillips and Land, 2012), abusing drugs (Bauld et al., 2010) and suffering physical and mental health problems (Herber et al., 2019; see Chapter 6).[2] When it comes to employment, we are primarily interested in *total labour demand* – or the total number of available jobs in society – alongside the *unemployment rate*. We are also interested in the *global jobs gap rate*, which is the percentage of the total adult population that has an unmet need for employment (International Labour Organization, 2023). This includes all unemployed people, as well as individuals, particularly women in lower-income countries, who are outside the labour force due to factors such as unpaid care obligations and so do not appear in unemployment data (International Labour Organization, 2023). Global unemployment is currently close to a record historical low, but it is starting to increase in several countries, partly due to the effects of higher interest rates. In higher-income countries, low unemployment may provide a rationale for increasing AI use. In lower-income countries, the biggest challenge is the lack of high-quality jobs pushing many people into less beneficial informal work, and it is unclear how AI will affect this (International Labour Organization, 2023).

Most research on AI and employment breaks down jobs into bundles of *tasks*, and it forecasts the extent to which AI will be able to perform them. Using this approach, Frey and Osborne (2013) estimated that 47% of jobs were at risk of automation by 'computer-controlled equipment', leading to multiple follow-on studies (e.g. Manyika and Sneader, 2018; Smit et al., 2020). These initial studies typically found that lower-income jobs, characterised by routine, physical tasks, for example in driving and manufacturing, were most at risk from AI. On the back of recent progress in large language models, more recent studies, such as Eloundou et al. (2023), point to another scenario, where roles that involve generating and manipulating information, and those which generally require higher levels of education, such as translators, survey researchers and tax advisers, are more exposed to AI. However, the authors do not take a view on whether this 'exposure' will be positive or negative for the employees.[3]

Empirical research is starting to shed light on this question of positive or negative impact of AI exposure on employees by studying the actual effects that AI has on labour demand after it is deployed in the workplace.[4] Handel (2022) and Albanesi et al. (2023) find little support for an acceleration in job losses for exposed occupations in US and European employment data, respectively. Indeed, Georgieff and Hyee (2021) find a *positive* link between AI exposure and employment growth. A key driver of these positive results was high-income employees with strong digital skills who likely had the capabilities and freedom to adapt their roles in response to AI. Acemoglu et al. (2022) find less positive results, with some evidence for reduced hiring for roles with greater AI exposure, but the effect sizes were too modest to draw strong conclusions.

Researchers have also surveyed organisations that have deployed AI applications to understand the resulting effects. In a recent survey of UK business leaders, Hunt et al. (2022) found that introducing AI was associated with both *destruction* and *creation* of jobs. More recently, in a study covering eight Organisation for Economic

[2] If unemployment persists, or an individual never achieves employment in the first instance, it can have a 'scarring' effect, as individuals drop out of the labour force, skills are forgotten and hiring discrimination by prospective employers increases (McQuaid, 2017).

[3] Acemoglu and Restrepo (2019) distinguish between displacement, augmentation and reinstatement effects. Displacement refers to a scenario where AI replaces human employees at performing certain tasks, leading to job losses; augmentation refers to a scenario where employees use AI to enhance their productivity; and reinstatement refers to a scenario where AI leads to new tasks being added to existing jobs or the packaging of new tasks into new jobs.

[4] Empirical analysis of the effects of AI on employment is difficult. For other technologies, like industrial robots, practitioners have access to longitudinal data sets (International Federation of Robotics, 2022) on the number of robots installed in different countries. The robots are also deployed in a narrow range of sectors and locations. This has allowed researchers to build models that tease out the resulting economic effects in a way that is not possible for AI. See, for example, Oxford Economics (2019), Petropoulos et al. (2018), Graetz and Michaels (2015), Acemoglu and Restrepo (2017) and Dauth et al. (2017). The findings differ but broadly support a view that industrial robots have displaced low-skill manual jobs, boosted organisational and national productivity, and created new high-skill jobs, with positive spillover effects on job creation in other sectors.

Co-operation and Development (OECD) countries and almost 100 finance and manufacturing organisations that had deployed AI applications, Milanez (2023) found that almost 80% reported no change in overall job quantities. Rather, most firms had invested in AI to improve product or service *quality*, so their headcounts did not change. Some AI applications did replace employees in performing certain tasks, but most of those employees were assigned new tasks. In other instances, the AI applications were insufficiently effective in increasing productivity or quality of service to have any impact on employment. Where job losses did occur, it was primarily via attrition rather than redundancies. Taken together, the evidence suggests that AI has *not yet had major negative impacts on aggregate labour demand*, or unemployment, although this may be because most AI applications have not yet been especially transformative and employment effects take time to manifest.

Job quality

The OECD defines *low-quality jobs* as those with: (1) low earnings; (2) a fragile sense of security; and (3) a poor working environment (OECD, 2016). The broader literature on job quality, including recurring employee surveys, also highlights a range of factors in the working environment that can positively or negatively affect well-being, for example by providing a sense of purpose or engendering feelings of contentment or stress. The relative importance of these factors can differ by person, and evolve over time, including, for example, working-time arrangements, workplace relationships, autonomy, employee voice and potentially many more aspects (US Bureau of Labor Statistics).

Early evidence suggests that the relationship between AI and job quality is not straightforward. When it comes to wages, Felten et al. (2019) find a small positive link between exposure to AI and wage growth, suggesting that AI deployments may make employees more productive and capable of earning a higher wage. However, other studies have found little evidence of any impact (Albanesi et al., 2023). Acemoglu and Johnson (2023) also claim that the adoption of AI in the workplace may enable employers to more strongly monitor and surveil their employees. This may mean that employers can make employees work harder without necessarily having to increase wages to the same extent as would otherwise be required (this point also illustrates the connections between job quality, productivity growth and inequality). Beyond wages, studies show a mixed picture, with employees reporting both positive and negative impacts of AI on job quality attributes, and with different employees reporting divergent views about the impact of some applications. For example, Gutelius and Theodore (2019) found that introducing AI-based robotics into warehouses helped to reduce the monotonous and physically strenuous activity of lifting heavy packages, but it also put pressure on employees to work faster, reduced human contact and led to an increase in perceived scrutiny. Relatedly, Milanez (2023) found that some employees felt that AI applications caused their work to become safer and more rewarding by, for example, reducing the number of repetitive interactions. However, others felt that they had less privacy, higher work-intensity and more stress (Ribeiro et al., 2023).

Productivity growth

Productivity describes the *efficiency* with which an economy uses *labour* and *capital*. We are primarily interested in growth in *total factor productivity* (TFP), which describes increases in output that are due to innovation, including technological progress. TFP is more difficult to measure than labour productivity, which is consequently more commonly used. Science and technology advances, and the innovative products, services and processes that result, are the only way to significantly boost productivity growth, and economic growth, in the long term. Economic growth, in turn, is central to maintaining and improving *standards of living*. For example, increased growth in China over the past 40 years has enabled 800 million people to move out of poverty (defined as income of less than USD 1.90 per day), accounting for three quarters of the global reduction over that time

period (The World Bank, 2022). Since the global financial crisis in 2007–2009, the world has witnessed a sharp slowdown and sustained stagnation in global productivity growth, including, most worryingly, in low-and middle-income countries (Dieppe, 2021).

Bearing this context in mind, AI could potentially make human employees and machinery (e.g. computers and software) more productive by boosting *efficiency* – doing things that were already being done, but faster or at a higher scale – or by boosting *innovation* – doing things differently or doing new things (Manyika and Spence, 2023). AI could directly suggest novel ideas or make human employees more efficient, and in turn free them up to work on more innovative ideas. Indeed, some believe that AI will lead to an explosion in new ideas, with a corresponding explosion in productivity growth and economic growth (Clancy and Besiroglu, 2023; see also Vollrath, 2023). However, AI could also hamper productivity growth. For example, the introduction of email led to quicker asynchronous communication, but it also prompted concerns about information overload, disruptions to deep work and spam, and debates continue over its net impact (Bulkley and Van Alstyne, 2008).

Most evaluations of AI's impact on productivity growth have been short experiments where individuals use AI tools to carry out discrete tasks, with generally positive results. Peng et al. (2023) tasked software developers with using GitHub Copilot to implement an HTTP server in JavaScript, and they completed the task almost 60% faster than the control group. Noy and Zhang (2023) assigned writing tasks to a group of college-educated professionals, and they found that access to ChatGPT reduced the time taken and increased the perceived quality.[5] Brynjolfsson et al. (2023) provided an AI conversational assistant to more than 5,000 customer service agents in a real workplace. The number of issues resolved, per hour, increased by 14%, on average.

The three studies all found evidence to suggest that the AI applications disproportionately helped lower-skilled or more novice employees within an occupation more than their higher-skilled counterparts. This suggests that AI assistants could help to ease or shorten the learning curve, and they may have a role in employee training and upskilling programmers. However, study results also suggest that not all employees will be helped by AI impacts, which may be more in the realm of incremental gains in efficiency rather than a sharp uptick in transformative innovation. The impacts will also differ by sector, job and task. For example, Jia et al. (2023) used an AI conversational agent to support telesales employees. They found that previously top-performing employees benefitted more, because they were better placed to use the assistant to experiment and were more creative when interacting with the AI assistant to answer customer questions.

A recent suggestion is that generative AI tools could double US productivity growth over 20 years – a huge potential impact (Brynjolfsson et al., 2023). Another study, by Goldman Sachs, estimates that generative AI could boost annual productivity growth by 1.5 percentage points over a 10-year period following mass adoption (Hatzius et al., 2023).

For now, there is little evidence in the productivity growth data that such impacts are occurring.[6] A key question is whether benefits to individuals will translate into economy-wide productivity growth. For example, it may be that AI's recommendations help individuals to become more creative, but only in similar ways, thus leading to reduced growth in creativity at the economy level (Doshi and Hauser, 2023). AI may also lead to significant productivity gains in certain sectors, but the resulting cost savings may be spent on other sectors, like education and health, that do not become more productive, the so-called Baumol effect, thus resulting in

[5] Explaining their results, Noy and Zhang (2023) find that ChatGPT could substitute for human tasks, such as rough drafting, to allow humans to focus more on idea generation and editing.

[6] Productivity gains from AI may take time to materialise. One potential positive scenario, outlined by Brynjolfsson et al. (2018), is a J-curve, where AI's initial effects on productivity growth are minimal, or even negative, as organisations redesign their business models, before leading to a subsequent surge.

little impact on aggregate productivity growth.[7]

Inequality

We focus here on *income* and *wealth* inequality *between* (place-based inequality) and *within* countries (class-based inequality).[8] Chancel et al. (2022) found in 2022 that the richest 10% of the world's population took home 52% of the total income, while the poorest half took home just 9%. The *wealth* disparity is even starker, with the richest 10% owning 76%, the top 1% owning 38% and the poorest half owning almost nothing (2%). When the authors analysed this data over the past century and beyond, several stories emerged, some of which are more positive than others. Positively, over the past 30 years, and particularly since 2000, inequality between countries, or *place-based inequality*, has declined. Less positively, inequality *within* many countries has increased over the past three decades, for example in the US, Brazil and India. Inequality also occurs based on parameters beyond income or wealth percentiles. For example, Chancel et al. (2022) note that women earn less than 35% of global income, with an increase of just 5% since 1990.

There is little empirical evidence regarding how AI has affected inequality. However, we can identify several pathways through which it is likely happening. The first reflects which employees are best able to draw on AI, to enhance their own productivity and wages, and which employees face the greatest displacement risk. Studies by Felten et al. (2019) and Albanesi et al. (2023) suggest that high-income occupations may be disproportionately benefitting from AI exposure to date. However, other studies suggest that AI assistants disproportionately benefit more junior and lower-skilled employees, so this could help to reduce inequality, although this will likely only occur if these employees have sufficient negotiation power to request higher salaries (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Peng et al., 2023).

Another route through which AI is affecting inequality is via the type, distribution and location of new jobs that it enables (Ben-Ishai et al., 2024). The majority of leading AI research labs, start-ups and enterprises are located in urban centres in high-income countries, so we can assume that AI is directly and indirectly enabling well-compensated 'frontier' jobs in these locations, such as engineers, product managers and lawyers, as well as lesser-paid 'wealth' roles, such as fitness instructors, that provide services to these high-income employees.[9] This may be exacerbating inequalities within countries, and between higher- and lower-income countries. Lee and Clarke (2019) estimate that for every 10 new high-tech jobs created in the UK, seven new service jobs were created, of which six were 'low-skilled'. Once accompanying increases in housing costs were considered, low-skilled employees' real wages fell, thus exacerbating inequality.

The development of large AI models is also enabling, potentially millions, of new 'data enrichment' jobs (Kässi et al., 2021). Activists and academics have criticised the precarious working conditions and low compensation, which has led to the creation of recommended best practices for AI labs (Jindal, 2022), including the payment of a living wage (Graham et al., 2017; Gray and Suri, 2019). If conditions improve, it is possible that data

[7] The Baumol effect appears to be one reason for limited productivity growth in recent years (Kling, 2018).

[8] Efforts to articulate how inequality affects societal well-being are plagued by definitional and contextual challenges, and the poor reproducibility of studies (Ngamaba et al., 2018). Potential effects include worse rates of subjective well-being (Oishi et al., 2011) and intergenerational inequality (Corak, 2013), undermined access to health (Pickett and Wilkinson, 2015), education (OECD, 2018b) and public services, and reduced social trust and civic engagement (Schröder and Neumayr, 2023).

[9] Autor and Salomons (2019) find that, over the past century, technology has primarily enabled three types of new jobs: (1) 'Frontier' roles produce, install and maintain (and arguably also 'use') novel technologies, such as engineers who may help design new AI assistants; (2) 'Last-mile' roles carry out 'nearly automated' tasks that do not require high levels of technology-specific expertise, such as data enrichment workers who help to curate data sets for new AI assistants, but arguably there is a spectrum of expertise that may be created in these roles, with scope for domain experts to participate; and (3) 'Wealth' roles, such as personal trainers, cleaners and counsellors, provide luxury services to affluent employees in high-income technology jobs. This reality of high-income roles creating a larger number of service roles is not limited to technology, as Moretti (2013) and others have shown.

enrichment could help to reduce inequalities between countries, similar to the past effects of outsourcing on per capita income in India, India, the Philippines and Morocco. However, the history of science and technology, where advanced research and development (R&D) often leads to nearby start-up creation and longer-term agglomeration effects, suggests that an expansion in higher-quality data enrichment work will be no substitute – from a longer-term inequality perspective – to having frontier AI development and jobs in low- and middle-income countries, and in the lower-income regions of high-income countries (Gross and Sampat, 2022).

17.3. How Will AI Assistants Affect the Economy?

To our knowledge, no substantive research addresses the aggregate economic impact of AI *assistants*, as a specific class of AI, outside the studies on productivity growth discussed above. The impact will also depend on the type of AI assistants that are deployed and their specific characteristics (see Chapters 2 and 4). To illustrate these dynamics, we create a simple framework for assessing the potential economic impact of an AI education assistant and an AI programming assistant.

Effects on employment

We can pose a number of questions about the impact of advanced AI assistants on employment. Which occupations do we expect to be *directly affected* by these assistants? To what extent can employees in these occupations *adapt their tasks* or find *alternative jobs*? Can they use the assistant to *augment* their role? How do we expect consumer demand for their product or service to respond? To what extent do we expect entirely *new businesses* or *new production processes* and *jobs* to emerge? And do we expect a longer-term preference for humans vs capital (AI) in these production systems?

AI education assistant

Teachers and tutors will likely be one of the groups most affected by AI education assistants. Globally, there are approximately 85 million teachers (Ritchie et al., 2023). In the UK, past data suggests that there may be up to 1.5 million private tutors, with approximately 100,000 working full time – global data is lacking (Kirby, 2016). Some AI practitioners, such as Stuart Russell, have suggested that AI assistants, like ChatGPT, may lead to 'fewer teachers being employed – possibly even none' (Devlin, 2023), and some analysis of occupational exposure to generative AI identify teaching of some disciplines as highly exposed (Felten et al., 2023). However, there are many grounds to challenge this projection. First, historical evidence suggests that it will be difficult to successfully integrate AI assistants into formal education without strong involvement from teachers, a trend that could lead to a demand for *more* teachers rather than less. Second, the world is already facing a major teacher shortage, recently estimated at 69 million people by 2030 (UNESCO, 2022).[10] Third, even if more powerful AI tutors can help students to master knowledge and skills, a big if, they may be ill-equipped to provide other functions of education, including cultivating human-to-human skills, such as collaboration and listening, or the broader functions of socialisation and individuation.[11]

Over the longer term, if AI assistants were to displace teachers *en masse*, it would likely require a broader transformation of formal education, for example via a large shift to fully virtual schools, which already exist and

[10] The teacher shortage is also leading to teacher–pupil ratios in many countries that are far higher than experts recommend.

[11] Biesta (2009) identified three *functions* that education can perform. *Qualification*: Providing people with the knowledge, skills, understanding and judgement to *do something*, for example a specific job or general-purpose skills such as critical thinking. *Socialisation*: The transmission of norms, values, cultures, religions and traditions, with the primary goal of bringing individuals into existing orders and societal structures. *Individuation*: Enabling individuals to become more autonomous and independent in their thinking and actions.

educate approximately 350,000 students in the US (Irwin et al., 2023). However, leaving aside practical issues around student care and supervision, virtual schools in the US have so far shown subpar educational outcomes (Molnar et al., 2021). Such a scenario may be more likely in higher education, or adult learning, as students are more likely to have the self-discipline and motivation that evidence suggests is necessary to pursue 1-to-1 online learning. However, even leading virtual universities, such as the UK's Open University, generally target a *hybrid* offering, primarily because students desire in-person engagement and more collaborative learning (The Economist Intelligence Unit, 2020).

Private tutors may face a greater risk of displacement, as many AI assistants are explicitly modelled on imitating their services, and AI tutors could potentially be integrated into free or low-cost online platforms, such as those provided by Coursera and Khan Academy. However, many private tutors work part time, often via digital platforms, so they are arguably better placed than teachers to integrate AI assistants into their work or to adapt the tasks that they provide, such as providing tutoring on how to best use AI tools in various subject areas. Moreover, only a small percentage of students have access to a human tutor, and demand is growing, helped by governments starting to fund private tutoring for disadvantaged students (e.g. the UK's National Tutoring Programme (UK Government, 2023)). The primary challenge such programmes face is a lack of high-quality tutors. This may point to an opportunity for governments to evaluate and fund innovative tutoring organisations that combine human and AI tutors for disadvantaged students.

AI programming assistants

Software engineers, of whom there are approximately 27 million globally (Qubit Labs, 2022), will likely be the most directly affected by AI programming assistants. GitHub suggests that its Copilot tool is now 'writing' 30% of new code (Gain, 2021). Even if the figure rises to 80% or 90%, another big if, there are reasons to be confident that demand for human programmers will remain robust. First, demand for programming will remain strong due to rising AI deployment and the economy's broader *digital transformation*, which involves shifting products, services and processes online, where there is considerable scope to grow further.[12] Second, past studies suggest that programmers may be the employees who are best able to adapt their jobs and tasks to account for AI assistants, as many are at least somewhat self-taught and used to adapting to new technology (Vincent, 2023b).[13] Third, humans will likely remain central to ensuring that any AI-generated code is interpretable, secure and legally compliant. Finally, software is a fast-evolving industry, with new languages and techniques routinely emerging – this creates a dearth of data for training AI assistants on cutting-edge use cases.

Effects on job quality

The salient questions for assessing the impact of advanced AI assistants on employment relate, first, to what the *current state* of job quality is in the most affected occupations, and second to how AI assistants might affect wages, job security and other key drivers of job quality, such as job intensity and stress, autonomy, and employee relationships and collaboration.

[12]For example, ecommerce now accounts for almost 25% of retail sales in the UK, North America and much of Asia, but just 16% in Western Europe, 13% in Latin America and 3% in Africa, and there is considerable scope to grow further (Morgan Stanley, 2023).

[13]For example, a case study by Horton and Tambe (2020) showed how after Apple announced in 2010 that it would no longer support Adobe Flash, other Flash specialists, especially those who were younger, less specialised, or had good 'fallback' skills quickly transitioned away from Flash. Similarly, a study by Das et al. (2020b) of 170m US job postings from 2010–2018 shows how the tasks of 'IT jobs' evolved in response to new technologies, shifting towards machine learning, scripting languages and cloud solutions, and away from traditional software products and services that require workers to perform structured query language (SQL), Java, and data management.

AI education assistant

Many teachers face below-median wages and high degrees of stress, although the roles typically provide the opportunity to develop a strong sense of purpose, and private institutions and certain countries, like Singapore, provide better wages and working conditions. In the UK, almost half (48%) of teachers say their workload is 'unmanageable' (National Education Union, 2023), and 44% plan to quit by 2027 (Harrison, 2022). The primary challenges are workload and lack of supporting resources. In addition to a lack of teachers, these challenges are partly due to steady expansion in the domains that teachers are expected to cover, such as socioemotional learning and media literacy. Evidence on past technology deployment suggests that if educational institutions are mandated to integrate AI assistants, without accompanying training and resources, it could worsen job quality (Global Education Monitoring Report Team, UNESCO, 2023). However, teachers may also be able to use AI assistants to support them, for example by generating ideas, suggested adaptations or feedback on lesson plans and teaching materials, with some examples already starting to emerge (Wang and Demszky, 2023).

AI programming assistant

Globally, studies suggest that programmers are relatively satisfied with their roles compared to other professions, helped by the fact that wage trends are positive in much of the industry (Graziotin and Fagerholm, 2019).[14] Key job-quality challenges include time pressure, getting stuck when problem-solving, working with bad code or with poor coding processes, and information overload. Independent evaluations of the effects of AI assistants on programmer job quality are lacking. However, in a GitHub survey of Copilot users, 60–75% reported feeling more fulfilled with their job, less frustrated when coding and more able to 'stay in the flow' and to focus on more satisfying work when using Copilot, and 87% reported how it helped them to preserve mental effort during repetitive tasks (Kalliamvakou, 2022). One caveat is that, if programmer roles become more about interrogating AI-generated code, some may start to feel less connection to their end output, which can be an important driver of job quality (Bryce, 2018).

Effects on productivity growth

To assess the impact of advanced AI assistants on productivity growth, we need to address several questions. How will AI assistants affect *individual productivity* growth? To what extent will the assistant enable higher *efficiency* vs *creativity* and innovation? To what extent can we expect productivity changes for individuals to be mirrored at the *industry* and *economy* levels?

AI education assistant

Education assistants could increase human capital, that is the knowledge, skills and personal characteristics that make people productive across all sectors of the economy (Égert et al., 2022). For this to be substantive, AI assistants would have to directly or indirectly enable a large number of people to better cultivate a broad range of knowledge, skills and characteristics, including how to best use AI. Such a shift is badly needed, as progress in human capital is stagnating, as evidenced (imperfectly) by a lack of progress in the OECD's Programme for International Student Assessment of 15-year olds on reading, maths and science (OECD, 2023) and the Programme for the International Assessment of Adult Competencies in literacy, while progress in AI is accelerating (OECD, 2019). While *access* to education remains a critical issue for many people, the poor quality of education is arguably a bigger issue from a productivity growth perspective (Égert et al., 2022). AI

[14] In the US, the average wage of software developers is more than the average of all occupations (BSA, 2016).

assistants could *potentially* help to improve the quality of education and help students to use AI, thus leading to a significant boost in human capital and productivity, although such effects would not be quick to materialise.

AI programming assistant

The deployment of computing in the 1970s and 1980s initially had minimal impact on productivity growth, but it began to contribute more meaningfully during 1995–2005, before productivity growth began stagnating again. Some have suggested that the recent stagnation shows that digital technology is less transformative than technologies of centuries past (Cowen and Southwood, 2019; Gordon, 2017). Andrews et al. (2016) suggest that the issue is more to do with laggard firms and sectors. Peng et al. (2023) provide early evidence that programming assistants can boost individual efficiency (see also Vincent, 2023b). Their ultimate impact on productivity growth may depend on two factors. Will they be able to provide novel code suggestions that programmers would not otherwise be able to produce? Will programming assistants be able to enable more rapid digitalisation in organisations and sectors that have traditionally lagged, such as healthcare, social care and the civil service?

Effects on inequality

The salient questions regarding the impact of advanced AI assistants on inequality relate to which groups may disproportionately *benefit* or *suffer* as a consequence of this technology being deployed at scale, where this question can be assessed both *between* and *within* countries, and in relation to job creation, income and wealth (see Chapter 15).

AI education assistant

Significant inequities exist between and within countries with respect to access to quality education that both reflect and contribute to income and wealth inequalities. AI assistants may have the *potential* to help to reduce these. For example, UNESCO's Global Education Coalition aims to use online resources to bring education to students not currently attending school (UNESCO, 2023). However, historically, the most common result of introducing digital technology into education has been what Burns (2021) refers to as a *caste system*, in which 'the wealthiest students get online learning, poorer students get radio or TV … and the poorest students get nothing'. This is primarily due to the supporting infrastructure, resources, family use and teachers that are needed to make technology use successful (see Chapter 15). To the extent that teachers may be affected by AI assistants, for example in terms of job quality, these effects will fall disproportionately on women, who make up the majority of teachers worldwide, especially for younger students (European Parliamentary Research Service, 2020). If AI assistants help to boost access to education for those who lack it, this would disproportionately help girls, although the *gender parity* index for accessing and staying in education differs by country (UNICEF, 2022).

AI programming assistant

If programming assistants lead to higher productivity, wages and new frontier roles for software developers, this may exacerbate inequalities within countries, as the average wage for these roles is typically already above the median. However, initial studies suggest that AI assistants may help to make programming-based roles more accessible. For example, one study found that respondents who were 'learning to code' were more likely to use AI tools than 'professional developers' (Stack Overflow, 2023), while Peng et al. (2023) suggest that AI programming assistants could help people to transition into software development careers, pointing to the

potential training and upskilling opportunities that governments, AI labs and civil society could support.[15]

17.4. Policy Implications

Acemoglu and Johnson (2023), Brynjolfsson (2022) and Manyika and Spence (2023) argue that technological vision and policy changes are required if the benefits of new technologies are to be widely shared (see also Ben-Ishai et al., 2024). In this spirit, policymakers and technology companies could potentially take actions beyond the status quo to ensure more positive societal outcomes from the use of advanced AI assistants. Some economists are sceptical about the ultimate potential for such directed technological change, because anticipating the economic impact of technology is difficult, and policy interventions to steer technology could create unintended outcomes (Agrawal et al., 2023). Despite that, some options and policy choices do exist. In this section, we outline three main actions that policymakers, AI labs and other stakeholders can take to try to ensure positive economic outcomes for society.

Monitoring how AI assistants are affecting the economy

Unlike other technologies such as industrial robotics, we have no clear picture of how many AI assistants are being deployed across the economy or what sectors they are being deployed into, thus inhibiting our ability to understand their economic impact. Given the diffuse nature of AI assistants and the many ways that organisations could adopt them, there is no easy solution to hand, but policymakers, industry associations and others could explore:

- **Collecting empirical evidence:** Fund more studies to understand the effect that AI assistants are having, including on neglected areas like job quality (see Chapter 19).
- **Develop new monitoring techniques:** Explore new methods for using AI and disparate data sources, including occupation data, start-up funding data and cloud computing services data to build more timely assessments of how AI assistants are affecting the economy.

Shaping the type of AI assistants that get developed or deployed

Policymakers need to develop well-calibrated expectations about their ability to shape the evolution of AI assistants in socially beneficial ways. Labour-displacing AI systems could generate significant negative externalities and lock society into economic trajectories with fewer jobs, reduced job quality and more inequality than is socially desirable, thus potentially limiting future progress and warranting interventions to steer their development in human-augmenting directions (Acemoglu and Johnson, 2023; Brynjolfsson, 2022; Korinek and Stiglitz, 2018). However, overzealous attempts to direct the progression of AI could inadvertently foster protectionism and potentially lead to misguided efforts to centrally control or guarantee job availability. Such measures could generate unintended negative effects, such as stifling innovation, creating market inefficiencies and inadvertently perpetuating jobs or tasks that may have become obsolete. However, there are steps that policymakers, by working with AI labs, civil society and industry associations, could take, including:

- **Align on beneficial use cases:** Align on priority AI assistants to support, such as scientific research assistants and assistants for teachers or job seekers.

[15] Similarly, Tu et al. (2023) suggest that AI assistants may transform the work of data science from hands-on coding, data-wrangling and conducting standard analyses to assessing and managing analyses performed by automated AIs.

- **Develop a research agenda:** To support these beneficial use cases, work with industry and civil society to align on key research questions, develop supporting data sets and fund supporting R&D, including neglected foundational research.
- **Explore how to encourage the development and deployment of beneficial types of assistants**. This could involve public funding for RD, responsible adoption in the public sector and other measures (Acemoglu et al., 2020).

Shaping the impact of AI assistant deployment

To ensure that the effect of AI assistants on employees and society is as positive as possible, policymakers, AI labs, civil society and industry associations could explore:

- **Education and training** by carefully integrating AI, and AI assistants, into education and upskilling programmes, while not avoiding the non-AI-related challenges that have often limited the efficacy of such programmes on core education and employment outcomes.
- **Employee consultation rules** by upgrading existing approaches to employee consultation with respect to new technologies and drawing on best practice.
- **Broad-based policies to enhance economic resilience** by going beyond targeted policies to ensure that the economic impact of advanced AI assistants is economically beneficial, there are many other broad-based policy interventions that could help to mitigate negative impacts. These include interventions to strengthen the social safety net, active labour market policy interventions and industrial policies to spur economic growth that compensate for economic displacement and jobs losses (Juhász et al., 2023).

17.5. Conclusion

This chapter examined the potential economic impacts of advanced AI assistants, particularly with respect to employment, job quality, productivity growth and inequality. While there is currently little substantive research on the economic impacts of AI assistants, as opposed to the economic impacts of AI in general, it is plausible that advanced AI assistants will have significant implications for each of the four variables examined. To that end, we recommend further research into how advanced AI assistants may impact employment, job quality, productivity growth and inequality. We also encourage policymakers to adopt active measures for monitoring and understanding how AI assistants are affecting the economy, to explore different levers that can be used to shape their design and deployment, and to implement plausible and evidence-based policy interventions to promote socially beneficial outcomes for AI assistants.

Chapter 18

Environmental Impact

Juan Mateos-Garcia, Sims Witherspoon, Iason Gabriel

Synopsis: As positive use-cases for advanced AI assistants continue to emerge in support of climate action, there is significant uncertainty about their overall environmental impact. While the study of AI's energy consumption and carbon emissions is still taking shape, some factors suggest that AI assistants could lead to increased computational impacts. However, there are many opportunities to increase the *efficiency* of these processes and make them more reliant on carbon free energy. Ensuring that AI assistants have a net positive effect on the environment will require model developers, users and infrastructure providers to be *transparent* about the carbon emissions they generate, adopt *compute-* and *energy-efficient* techniques, and embrace a *green mindset* that puts environmental considerations at the heart of their work. Policymakers may also want to create incentives that support these changes, minimise the environmental impact of AI systems deployed in the public sector, support AI applications to tackle climate change and improve the evidence base about the environmental impacts of AI. Promisingly, it may be possible to develop AI assistants that broaden access to environmental education and scientific evidence – and that improve the productivity of engineering efforts for climate action.

18.1. Introduction

Anthropogenic climate change, driven by greenhouse gas emissions, is one of the most pressing issues facing our planet today. The effects of climate change are already being felt around the world in the form of more frequent extreme weather events, rising sea levels and changes in plant and animal life (Calvin et al., 2023). The development of large foundation models and widespread adoption of AI services could potentially contribute to these impacts (Bender et al., 2021). At the same time, with sufficient attention, technical and algorithmic innovations, better infrastructure, and access to carbon-free energy, we may be able to contain the potential environmental impact of AI or even reverse it over time, by improving productive efficiency and enabling innovations that contribute to wider environmental sustainability (Dannouni et al., 2023; Patterson et al., 2021; Rolnick et al., 2019).

Promising applications for AI in this space include efforts to produce carbon-free and low-carbon electricity via better forecasting (Lam et al., 2023) and scheduling of energy supply and demand for renewables, better storage technologies, and support to new sources of energy such as nuclear fusion (Degrave et al., 2022). AI technology may also help to reduce the impact of transportation systems by modelling demand, improving freight routing and encouraging the adoption of electric vehicles. Lastly, AI may help to reduce greenhouse emissions from buildings and cities via developments in the field of smart buildings and smart cities (Luo et al., 2022). These trends create substantial uncertainty about the overall direction of the impact of AI on carbon

emissions.

This chapter considers the question of environmental impact in the context of increasingly advanced AI assistants. What might their environmental impact be? Can they contribute to sustainability efforts and, if so, how? Lastly, what techniques, actions and policies can be used to steer their development and deployment towards sustainable outcomes? Although we focus primarily on the link between AI and greenhouse gas emissions, it is also important to note that the development of powerful AI systems has a wider environmental impact – including on global water consumption, the mining of minerals and the generation of toxic emissions (Crawford, 2021; Dannouni et al., 2023; Li et al., 2023b). Much of the discussion in this chapter is relevant for mitigating those other impacts as well.

18.2. The Environmental Impact of AI Systems

Climate change

In its latest assessment report, the Intergovernmental Panel on Climate Change (IPCC) shows, with high confidence, that human-caused climate change is affecting communities across the globe (Calvin et al., 2023). This includes rising sea levels, extreme weather events such as heatwaves and droughts, mass species extinction and impacts on agriculture and fishing (Calvin et al., 2023). Moreover, the IPCC estimates that with only current mitigation efforts in place, the goals set out in the Paris Agreement are likely to be missed, with the global temperature rise exceeding 1.5°C in the first half of the century, potentially even reaching 2°C – which is a threshold that could have catastrophic consequences.

By way of illustration, the World Health Organization estimates that climate change could cause an additional 250,000 deaths annually between 2030 and 2050 (World Health Organization, 2018), while global gross domestic product (GDP) could also decline by an average of 10% (Swiss Re Institute, 2021). It has also been estimated that 143 million people could be forced to migrate in response to climate emergencies by 2050, with communities in low- and middle-income countries particularly affected (Rigaud et al., 2018).[1] These impacts would likely be accompanied by significant societal disruption, and by economic and political instability, posing a significant threat to human rights and global security (Levy and Patz, 2015).

To achieve the goals set out in the Paris Agreement, the world needs to reduce global CO_2 emissions by 45% by 2030 and to reach net zero by 2050, but the world is not on track to meet those targets (Ritchie et al., 2020; UNFCCC. Secretariat, 2023). Action is therefore needed across a range of fronts. At a general level, energy consumption (including transportation, electricity, heat, building construction, and manufacturing) contribute the majority of carbon emissions worldwide (75.6%), followed by agriculture (11.6%), industrial processes (6.1%), waste treatment (3.3%) and land use (3.3%) (Ge et al., 2020). Against this backdrop, cloud services and large scale data centres – where the majority of AI computation takes place – account for 0.1-0.2% of greenhouse gas emissions, with around 25% of their traffic related to AI (Kaack et al., 2022).

Nonetheless, given the wider context it is important to understand and mitigate the potential environmental impacts of powerful general-purpose AI systems – and to maximise any contribution they can make to efforts to tackle climate change. Here, we draw on a framework developed by Kaack et al. (2022) to distinguish between *computational impacts*, *application impacts* and *systemic impacts* of computer systems. We begin with a broad assessment of AI's environmental impacts in this section before focusing on assistant-specific issues in the next one (see Table 18.1 for a summary).

[1] These impacts are particularly felt in low- and middle-income countries that have historically created low levels of CO_2 emissions – raising further questions about economic and environmental justice (Calvin et al., 2023; Wenz, 1988).

Computational impacts

Data and compute are basic inputs into the development and deployment of modern AI systems (see Chapter 3). Accessing these resources requires dedicated hardware for storage and processing (such as graphics processing units (GPUs)) and infrastructure including data centres and telecommunication networks. Interacting with AI systems also requires user and consumer hardware such as smartphones (to access AI assistants) or gaming consoles (to access AI-enabled video games), all of which require energy to function.

This hardware creates two types of environmental impact: first, there are *embodied impacts* created through its material collection, manufacturing and delivery. For example, the creation of semiconductors requires extraction of raw materials, manufacturing using large amounts of energy, water and hazardous chemicals, and emissions-producing transport to the delivery destination (Kuo et al., 2022). Second, *operational impacts* are created when an AI system is designed, trained and deployed. Indeed, AI systems continue to consume energy after they are deployed whenever they make inferences, for example, in response to user queries. In addition, their availability to respond to user queries incurs idle energy use when they are not actively running inferences (Luccioni et al., 2022).

A growing body of literature has started to study AI's energy consumption and carbon emissions (and its drivers) using a variety of methodologies. Some key themes include:

- **Variation in estimates about how energy intensive modern AI systems are likely to be**. In their analysis of the environmental impact of machine learning (ML), Strubell et al. (2019) estimated that training a single transformer model with neural architecture search generates CO_2 emissions that are equivalent to the operation of five cars during their lifetime (see also (Kuo et al., 2022; Strubell et al., 2020). However, this estimate is challenged by Patterson et al. (2022) who argue that the impact is several orders of magnitude less. More recently, Luccioni et al. (2022) compiled information about state-of-the-art models to show that, for example, training GPT-3 created 552 tonnes of CO_2 equivalent, which is the same as that used for 550 round-trip flights between New York and San Francisco (see also Stokel-Walker, 2023). Additionally, Luccioni and Hernandez-Garcia (2023) estimate that carbon emissions are growing over time for a sample of AI systems (mostly developed in academia), with question–answering systems experiencing the fastest overall growth in emissions. Most analysis of the environmental impact of AI focuses on emissions during training, because it is relatively easy to measure how much energy is used in this way. As an exception, Luccioni et al. (2022) estimate that the embodied emissions of BLOOM, a 176-billion parameter open-science, open-access language model developed in 2021–2022, account for 22% of its total CO_2 emissions.
- **Inference may be more important than training when it comes to energy use.** Luccioni et al. (2022) also estimate the emissions generated during the deployment of BLOOM, calculating that the model emitted 19 kg of CO_2 emissions per day throughout the monitoring period. In their survey of the environmental impacts of AI systems at Facebook, Wu et al. (2022) note that some use cases – such as language models – generate two thirds of their operational emissions during inference, while in other cases the carbon footprint is more evenly distributed between training and inference (something that depends on the size of the model, how many users it has and how often it has to be retrained). Focusing on overall energy usage, Patterson et al. (2022) found that 60% of ML energy use at Google from 2019-2021 was attributable to inference.[2]
- **Demand for larger models and improvements in operational efficiency have countervailing effects on energy consumption and emissions**. Scaling laws suggest that increases in model size (which require

[2]Note that the actual emissions generated by this energy usage depend on the grid electricity mix.

larger amounts of compute for training and inference) are associated with predictable improvements in model capabilities (Kaplan et al., 2020). This is linked to rapid growth (a doubling every 5–6 months) in the amounts of compute used to train state-of-the-art AI models, which could increase energy demands (Sevilla et al., 2022). At the same time, Patterson et al. (2022) argue that improvements in model efficiency and infrastructure will ultimately help contain the energy costs and environmental impacts from larger models. They compare training GPT-3 with GLaM, with the latter more recent model becoming 2.8 times more efficient as a result of model improvements. Furthermore, running GLaM on low-carbon infrastructure greatly reduced the estimated resulting carbon emissions. We discuss various technical and infrastructural levers, which can be used to mitigate the environmental impacts of AI and AI assistants in more detail in Section 18.4.

- **Environmental impacts are shaped by factors internal and external to the development process**. When looking at internal factors, model size seems to be the main determinant of energy costs. Larger models generally require longer training times using more energy-intensive GPUs. On this point, Luccioni and Hernandez-Garcia (2023) find a strong correlation between training time, energy consumption and CO_2 emissions. When looking at external factors, data-centre energy efficiency and the carbon intensity of the electricity grid – which may depend on high-emission energy sources vs low-carbon sources – are a key determinant of a model's CO_2 emissions (Dodge et al., 2022; Wu et al., 2022). This points at the possibility of decoupling AI energy consumption from carbon emissions through the use of carbon-free energy sources, a point we return to in section 18.4.
- **Analyses of the environmental impacts of AI rarely consider counterfactual scenarios**. AI systems are deployed to undertake economically valuable activities which, in their absence, might have to be performed using alternative production processes and technologies that also consume energy and generate carbon emissions. These counterfactual or 'replacement' costs are generally neglected in the literature but remain a key part of determining overall impact.

Application impacts

AI systems also have an impact on the environment through the applications that they enable. Kaack et al. (2022) distinguish between applications that contribute to environmental sustainability and applications that increase carbon emissions, thereby accelerating climate change.

Rolnick et al. (2019) explore the first type of application – those that have the potential to *mitigate* climate harm – across thirteen areas, ranging from electricity systems to collective decision-making. In broad terms, AI systems can help to tackle climate change by improving our understanding of its extent, drivers and impacts, by optimising systems to mitigate those impacts and by accelerating the transition to a sustainable economy. Dannouni et al. (2023) distinguish between three use cases for AI in climate action – mitigation (including measurement, monitoring, reduction and removal), adaptation and resilience (including hazard prediction and vulnerability management), and foundational capabilities (for climate and economic modelling, behavioural change, and innovations and breakthroughs). Salient examples include AI systems that can be used to predict extreme weather events (Ravuri et al., 2021), reduce the energy consumption of industrial cooling systems (Wong et al., 2022), help design plastic-eating enzymes (Kincannon et al., 2022) and accelerate fusion science (Degrave et al., 2022).

On the flipside, AI could potentially increase the productivity of extractive and carbon-intensive industries such as oil and gas or cattle farming, helping them continue or even ramp-up their activities (Greenpeace, 2020; Kaack et al., 2022).

Systemic impacts

The last category in the (Kaack et al., 2022) framework aims to capture the systemic impacts of AI which encompass indirect and second-order environmental effects. This includes situations where AI-driven improvements in efficiency or costs lead to a rebound effect in production and CO_2 emissions (referred to as 'Jevon's paradox') – or cases where AI entrenches or displaces unsustainable technologies and consumption patterns. For example, more capable self-driving cars could reduce the use of public transport (Lanzetti et al., 2021), which would be potentially problematic from a climate perspective. Alternatively, AI services could reduce demand for travel via better scheduling and more effective and immersive video-conferencing. These dynamics, though important, are often speculative and tend to be particularly hard to model or evaluate accurately (see Chapter 19).

18.3. The Environmental Impact of Advanced AI Assistants

Having set out a general framework for understanding the environmental impacts of AI, we consider here the special case of advanced AI assistants. How might the nature of their development and capabilities shape those impacts?

AI assistant computational impacts

Three features of AI assistants suggest possibly increased computational impact.

First, AI assistants are likely to be based on transformers trained on GPUs during long runs in some cases involving multimodal learning (see Chapter 3), which can be more computationally intensive (Xu et al., 2023). Looking at foundation models, GPT-3' training is estimated to have generated 552 tCO_2 emissions. More recently, Meta's Llama 2 – a collection of open models optimised for dialogue – emitted 539 tCO_2 during training (Touvron et al., 2023).[3] All this suggests that the development of more advanced AI assistants, animated by powerful foundation models, could be an energy and carbon-intensive technology activity. At the same time there is scope to significantly mitigate these effects through improvements in efficiency that reduce energy consumption and through use of carbon-free sources that break the link between energy consumption and carbon emissions.

Second, AI assistants are targeted at consumer markets, suggesting potential operational impact through inference (see Chapter 3). As Wu et al. (2022) note, this kind of large-scale deployment produces the bulk of emissions for LLMs. For example, the BLOOM model tested by Luccioni et al. (2022) generated 340 kg of CO_2 emissions after receiving 230,000 queries during two weeks of deployment. AI assistants deployed in large consumer markets over longer time-frames could have bigger impacts.

Third, there are additional environmental impacts from downstream AI development activities such as the use of reinforcement learning from human feedback (RLHF) to improve model usability or the use of tools and APIs by AI assistants to enhance assistant capabilities (Ziegler et al., 2020; Schick et al., 2023; see Chapter 4). This latter class of impacts could be particularly significant if advanced AI assistants make substantial use of other AI tools – such as question answering systems or programming tools – that consume additional energy.

Further areas of uncertainty that have the potential to significantly affect the calculus discussed so far include:

[3] It is worth noting that this includes several models of different sizes. The largest Llama 2 model (70B) produced 239 tonnes of emissions. The model developers point out that they offset all emissions generated by Llama 2 through Meta's sustainability program.

Table 18.1 | AI and AI assistant environmental impacts

Type of AI impact	Factors relevant for advanced AI assistants
Computational Impacts *embodied* in production and distribution infrastructure, including construction processes, materials and resources. *Operational* impacts during exploration, training and deployment (including inference).	*Drivers that might increase emissions* • Size of underlying models • Scope for mass-market deployment • Additional impacts during development (e.g. Reinforcement Learning from Human Feedback) and deployment (e.g. tool use by AI assistants) *Drivers that might decrease emissions*: • Efficiency measures during model development and training • Improving access to carbon-free energy options *Areas of uncertainty* • Scale of embodied emissions • Role of smaller models and edge computing
Application The use of AI for applications with *direct* impacts on the environment (either positive or negative).	*Applications of assistants with environmental implications* These impacts are expected to be *moderate* in the near term because: • Sustainability-related use cases in education, science policy, software development and research and development (R&D) generally create indirect / long-term positive impacts through coding, scientific R&D and education. • AI assistant use cases for the extractive industries are likely quite limited.
Systemic Indirect environmental impacts created by AI (e.g. changes in consumption patterns)	*Areas of uncertainty* Systemic impacts could be important but are difficult to estimate with any degree of precision. Beneficial impacts could result from improvements in public awareness and education about climate change. Conversely, the potential contribution to increased general consumption or environmental misinformation are sources of risk.

- **Scope for mitigating impacts via efficiencies in model development and deployment, and by sourcing energy from low-carbon sources**. As noted previously, the development of energy-intensive AI systems has been accompanied by parallel improvements in the efficiency of processes and infrastructure used for AI model development and training. There is also significant scope to further shift energy consumption towards low-carbon sources that generate less emissions. We discuss these considerations next section.
- **Scope for reducing emissions by efficiently training only a few foundation models.** These models could then be fine-tuned downstream instead of training many separate foundation models *de novo*.
- **Demand for smaller models.** There are instances where small fine-tuned models, such as Alpaca, perform competitively when compared to larger models. If smaller models prove popular, they could be deployed with much lower operational impacts. These models could also potentially be run on 'edge devices' such as smartphones, thus increasing their energy efficiency and reducing data transport costs (Qualcomm, 2023).
- **The scale of embodied emissions.** As we noted above, there is a dearth of data about the embodied impacts of AI on the environment, including production, transport and disposal. However, it has been argued that these effects could be substantial (Crawford, 2021; Gupta et al., 2020; Patterson et al., 2021). If assistants increase the demand for AI services and related hardware, including semiconductors and data collection and storage, this could intensify their overall impact.
- **Counterfactual energy costs and emissions generated by providing assistant-like services without AI.** Advanced AI assistants are likely to provide economically valuable services to their users which, in their absence, would have required alternative processes and technologies that also consume energy and generate emissions. It is important to take these into account when assessing the environmental impacts of AI assistants.

Taken together, these dynamics provide reasons for vigilance when considering the environmental impacts of advanced AI assistants and suggest that choices about how advanced AI assistants are designed and operationalised are likely to be consequential from an environmental perspective.

AI assistant application impacts

Rolnick et al. (2019) classify AI technologies based on their relevance for 36 solution domains that may help to tackle climate change. The closest technology to AI assistants that they consider is natural language processing (NLP) (see Chapter 3). Their assessment suggests that NLP is unlikely to be transformative for efforts to tackle climate change: it is highlighted as relevant for seven solution domains (i.e. 20% of the total) and is expected to generate indirect and generally longer-term impacts. However, it is also important to note that their assessment was based on NLP-capabilities circa 2019. More advanced AI assistants could be expected to create new opportunities for developing environmental applications either directly or indirectly. For example, they might improve the efficiency of scientific R&D overall, improve software development for environmental use cases (Peng et al., 2023), or help to synthesise scientific evidence to inform policy (Tyler et al., 2023). The initiative Climate Policy Radar is a good example of this last effect.

One important solution domain where AI assistants could help is education, where Ge et al. (2020) and Rolnick et al. (2019) suggest that AI-powered tutoring systems could democratise and improve the public understanding of climate change via personalised and translated content (e.g. simulating the impact of climate change on a learner's location) (Amini et al., 2023). In the service of these educational goals, large language models such as ClimateBERT, and chatbots like ChatClimate, have been trained on high quality data from

authoritative sources – in order to provide access to trustworthy information about environmental science and climate change (Vaghefi et al., 2023; Webersinke et al., 2021).[4] Nonetheless, it is worth noting that although there is evidence that educational interventions can contribute to pro-environmental behaviours, the benefits are indirect and more likely to materialise in the the long term (Begum et al., 2021; see also risks arising from Chapters 9 and 16).

AI assistant systemic impacts

The indirect nature of system-level impacts makes analysis of them both speculative and hard to reliably measure (see Chapter 19). Some examples, in the case of AI assistants, include the potential for shifts in demand and efficiency in energy-intensive activities such as computer programming (which are made cheaper by AI-coding tools) and potential shifts in overall levels of misinformation or polarisation (which can impact society's collective ability to respond to climate change, see Chapter 16).

18.4. Mitigating Negative Environmental Impact

Having outlined the factors that might drive or help to mitigate the environmental impacts of AI assistants, application opportunities and systemic effects, we now consider the implications for model developers and adopters, primarily in industry and government, who are committed to helping to achieve international climate goals (also see Dannouni et al. (2023)).

Levers for model developers and users

Model developers and users need to prioritise the adoption of technical measures that mitigate and reduce the environmental impacts of AI assistants. Such measures are needed both ahead of time and throughout their deployment. In the context of AI energy efficiency at Google, Patterson et al. (2022) argue that better model architectures could help contain the environmental impact of this technology. They note, for example, that sparse models require five to ten times less compute than those following more denser model architectures. Furthermore, Wu et al. (2022) suggest other strategies that may prove helpful in reducing computational costs. These include careful data scaling, selection and sampling, and developing memory and data-efficient architectures. Model developers and users should adopt these and other techniques with the goal of mitigating the environmental impacts of AI systems.

Model developers and users should use hardware and infrastructure that minimise the environmental impacts of AI systems. Several factors, beyond AI model design and training, influence the overall energy consumption and emissions produced by the AI ecosystem. This includes the energy efficiency of hardware and data centres and the carbon intensity of the electricity grid. Patterson et al. (2022) identify opportunities to reduce energy consumption one-hundred fold, and carbon emissions by up to one thousand times, by adopting sparse architectures, using processors optimised for ML training, energy-efficient cloud services, and by optimising the allocation of computing workloads across data centres to maximise the use of clean energy sources (see also Schwartz et al. (2020)).

Model developers and users should, when possible, source carbon-free energy for the data centres where AI systems are trained and inferred. Drilling down further, into the AI carbon emissions produced by cloud computing, Dodge et al. (2022) note that the choice of region and time of day, and the use of workload optimisation methods that take into account carbon emissions can play a substantial role in mitigating

[4] ClimateBert provides a detailed emissions scorecard including grid emissions details in its website.

AI's environmental impacts. Model developers and users can harness these opportunities to minimise the environmental impacts of AI systems.

Model developers and users need to support transparency around the computational efficiency and energy consumption of AI models and infrastructure. Mitigating the environmental impacts of AI assistants is likely to require better data about those impacts. To support this goal, AI models' computational efficiency and energy costs can be incorporated into development and benchmarking evaluations, thus reducing the need for *ex post* estimations of energy costs or emissions based on incomplete information (Patterson et al., 2022; see Chapter 19).

A growing number of initiatives have appeared in this space, including the creation of various tools to measure model carbon emissions (summarised in Luccioni and Hernandez-Garcia (2023)), the adoption of new model evaluations (see Luccioni and Hernandez-Garcia (2023); Solaiman et al. (2023)) and changes in submission information requirements for major computer science conferences such as NeurIPS and NAACL. This information could then be used to implement labelling schemes that increase the visibility of model developers and cloud providers adopting good sustainability practices – building on the growing adoption of techniques to document model development and risks such as model cards (Mitchell et al., 2019). However, the proliferation of transparency requirements and voluntary reporting initiatives should be carefully coordinated, as they could otherwise create duplication and inconsistencies in reporting – and increase the overall reporting burden, without creating the desired benefit.

Model developers and users need to adopt a 'green AI' mindset in relation to the environmental impact of their choices and work. Each of the aforementioned initiatives needs to be underpinned by a reorientation in the AI R&D community towards a 'green mindset' that foregrounds environmental considerations. By remaining attentive to these questions and maintaining awareness of the relationship between the climate crisis and their own work, developers can take proactive measures to mitigate the harmful environmental impacts of AI by treating this as an important goal – alongside other considerations such as model accuracy (Schwartz et al., 2020). The Green Software Foundation is a good example of an effort in that advances this perspective and way of thinking.

Policy levers

Policymakers may want to consider policies that improve access to low-carbon or carbon-free energy for model development and deployment. This includes accelerating grid decarbonisation with standards and incentives that help accelerate the deployment of carbon-free energy (including via the adoption of AI systems to optimise the grid), funding research and development for emerging carbon-free energy technologies that may enable the deep decarbonization of grids, as well as greater retail access for model developers and users to source carbon-free energy for their operations (Golin and Sweezy, 2022).

Policymakers should ensure that the public sector leads by example in the development and adoption of sustainable AI methods and adopts a green AI mindset. This includes funding research to develop sustainable AI methods, ensuring effective transparency around energy consumption and emissions, adopting sustainable model development and training strategies, and prioritising energy-efficient cloud infrastructure when conducting research into AI as part of a public project (Dannouni et al., 2023; Ho et al., 2021). These interventions would have the added benefit of increasing the supply and potential flow into industry of AI scientists and engineers with a green AI mindset and skill base.

Policymakers should explore how to enable the development of AI applications that contribute to sustainability. This includes direct funding for AI research with sustainability applications, increasing the

availability of data sets to train those applications, building expertise and supporting deployment in relevant sectors (Dannouni et al., 2023; Global Partnership on AI, 2021).

Policymakers can help to reduce uncertainty and improve the evidence base about AI's computational and systemic impacts on the environment. Measures of this kind include supporting the development of better methodologies to measure the environmental impact of AI (taking into account counterfactual impact from delivering products and services without AI support - the EU Green Digital Coalition is already driving progress in measuring the net environmental impact of digital technologies), supporting agencies that can implement these methodologies impartially – and encouraging transparency around energy costs and CO_2 emissions by model developers, users and infrastructure providers, in a way that is *standardised* and avoids duplication and excessive administrative burdens.

18.5. Conclusion

The climate crisis has put advanced AI systems in the spotlight, both as potential contributors to climate change and as a potential source of solutions for tackling it. In this chapter we draw connections between these matters and the development of advanced AI assistants. Our analysis points towards the existence of deep underlying uncertainty about the likely impact of advanced AI assistants on climate change. While the foundation models, upon which advanced AI assistants are based, continue to grow in size, a range of substantial and successful efforts – aimed at increasing operational efficiency and leveraging low-carbon or carbon-free sources during training and deployment processes – are now underway.

Model developers, users, infrastructure providers and policymakers all have a role to play in ensuring that the technical and non-technical measures needed to reduce the impact of AI systems are identified and realised, that there is transparency about these impacts, and that the sector continues to cultivate a green AI mindset – one that places as much emphasis on environmental considerations as on other dimensions of model performance (see Chapter 5). By acting in concert, these stakeholders can help ensure that advanced AI assistants have a net positive impact on the environment.

Chapter 19

Evaluation

Laura Weidinger, Maribeth Rauh, Lisa Anne Hendricks, Arianna Manzini, Nahema Marchal, A. Stevie Bergman, Geoff Keeling, Will Hawkins, Iason Gabriel, William Isaac

Synopsis: This chapter provides a high-level introduction to AI evaluation, with a specific focus on AI assistants. It explores the *purpose* of evaluation for AI systems, the *kinds* of evaluation that can be run and the *distribution of tasks* across three layers of output (the model level, user-interaction level and system level) and among different actors. The chapter notes that, with regard to many salient risks and goals that we need to attend to in the context of AI assistant development, there are significant evaluation shortfalls or gaps. To address these limitations, the chapter explores what a more complete suite of evaluations, nested within a robust evaluation ecosystem, would look like and makes recommendations on that basis.

19.1. Introduction

This chapter focuses on the evaluation of advanced AI assistants. As has become clear throughout this paper, AI assistants raise a range of ethical concerns and considerations, including concerns around value alignment, privacy, anthropomorphism, misinformation, and safety. How do we proceed from here? To assess, prioritise and address the ethics of AI assistants, potential risks and benefits must be understood as comprehensively as possible. Alongside exercising *foresight*, which is needed to identify potential risks, and *monitoring* real-world outcomes (including unexpected failure modes and accidents as a technology is deployed), *evaluation* is a key component of understanding potential risks and benefits.

Evaluation is the practice of assessing the capabilities, robustness and impacts of an AI system against goals or risks. For example, we may evaluate the likelihood that an AI assistant disseminates misinformation, or we can assess the impact that using an AI assistant has on people's well-being. Evaluation requires operationalising potential harms into tractable, measurable observations. It also requires making a normative assessment of what merits evaluation in the first place, and at what stage AI assistant performance is 'good', 'fair' or 'safe enough'. By providing information on AI assistant capabilities, robustness and impacts, evaluation can play two critical roles for the ethics of AI assistants: it can guide AI iterative model *development* by providing a target for AI developers, and it can provide *assurances* and inform responsible decision-making on the design, risk mitigation and release of AI assistants.

In this chapter, we describe the evaluation of AI assistants as a fundamental building block to creating ethical systems. We proceed as follows. We first provide an overview of the practice of evaluating AI systems. We then survey existing approaches for evaluating AI assistants for the risks of harm discussed in this paper, before discussing the limitations and gaps of such approaches. The conclusion summarises the chapter.

19.2. Evaluating AI Systems

Evaluation is the practice of assessing an AI system's *capabilities*, *robustness* or *impact*. Capabilities refer to the functionality and usability of the AI system, and its limitations. Can it perform the tasks that its developers intend? Does it function equally well in different languages, and is it accessible to differently abled users (see Chapter 15)? Robustness refers to predictability and reliability of the system (i.e. the degree to which the system will *consistently* behave in acceptable ways, including in novel situations). The impact of an AI system refers to its effects on people who directly or indirectly interact with it, and on broader structures in which the AI is embedded, such as the natural environment, society and the economy. Through evaluation, the capabilities, robustness and impacts of an AI assistant can be understood. This is particularly critical for AI assistants, given the ethical risks they pose. In this chapter, we outline the *aims of evaluation*, *elements and varieties of evaluation* and a *three-layered evaluation framework* to comprehensively assess ethical considerations on AI assistants.

Aims of evaluation

Evaluation serves two main aims. It is critical for guiding AI assistant *development* and informing *responsible decision-making* at multiple time points over the life cycle of an AI assistant.

Evaluation serves a critical function in AI *development*. AI developers perform some evaluations at regular intervals over the course of AI system development to track AI system performance against established tests. Performance on these tests is taken as an indicator of overall progress in AI model development, and it can be compared against performance of other AI systems. Surpassing other AI systems or reaching certain thresholds on these established tests is often read as an important signal that an AI system has reached or exceeded state-of-the-art levels of performance in the field. In this process, evaluation provides a regular signal that guides model development and is a fundamental building block of iterative design. This process is also referred to as 'hill climbing' (Sutton, 2019; see Chapter 3).

Second, evaluation is a critical component of *responsible decision-making* at different time points and for different actors, including AI developers, users of AI assistants, regulators and civil society. For AI developers, evaluation underpins decisions on guiding AI development, anticipating and mitigating potential risks, and whether work on an AI system should be stopped until certain concerns are resolved or harms are mitigated (Shevlane et al., 2023). Evaluation can help product developers to compare functionality for different possible user groups and underpin normative decisions such as whether an AI system's performance is 'safe enough' or 'good enough' for use in a given context (Bakalar et al., 2021; see Chapter 5). Evaluation can also inform potential downstream users of an AI assistant to understand in what contexts it is safe to use. For public authorities, civil society and users of AI assistants, evaluation results are required to ensure that an AI system is used in contexts in the real world for which it has been tested and is safe (Mitchell et al., 2019; see Chapter 7).

Elements and forms of evaluation

Evaluation consists of technical and normative steps. It requires (1) *selecting a target* for evaluation, such as a risk of harm or a performance goal to measure, (2) *operationalising the target* into a concrete *test* or *metric*, which may require trading off different considerations, and (3) *assessing results against* established *thresholds or aims*. A target in evaluating the ethics of AI assistants may be the likelihood of the AI assistant to mislead a user (see Chapter 16), or the factors that affect the extent to which users anthropomorphise the AI assistant (see Chapter 10). Operationalising the evaluation target into a concrete test and metric requires a *theory* of how a given harm can be detected and measured. There are often multiple ways in which a target can be

operationalised, and deciding how to define and measure the target is a normative and contestable decision. For example, misinformation risks may be operationalised as the frequency at which an AI assistant outputs correct vs incorrect statements (Lee et al., 2023b; Lin et al., 2022). It could also be operationalised as how likely users are to believe false outputs that are AI generated (Bai et al., 2023). Alternatively, it could be evaluated by measuring the broader spread of misinformation as there is an increasing uptake in AI assistants (Allen et al., 2020). Operationalisation may also require normative trade-offs, for example on how to weigh false positives against false negatives in a given metric. Once a measurement is obtained, the third piece of evaluation is to assess the observed model performance against a normative threshold or an aim. Note that evaluation is never neutral: it always requires a normative judgement on how different results should be valued (Bowker and Star, 2000).

Evaluation can take different forms. Most commonly, evaluations of AI systems are *automated performance tests* against tasks or data sets aimed at capturing capabilities of interest. These automated evaluations may target concepts as narrow as 'correctly identify a user's voice' or as broad as 'helpfulness'. Additional modes of evaluation that leverage human expertise, including human evaluations (Glaese et al., 2022; Thoppilan et al., 2022), adversarial testing (Perez et al., 2022b), user testing (Lee et al., 2023a), bespoke tests of AI assistants in particular contexts (Marda and Narayan, 2021) and expert assessments of AI impacts (Raji et al., 2020b), are increasingly being developed and are indispensable for a comprehensive evaluation of the ethics of AI assistants. Pre-deployment evaluation is complemented by post-deployment monitoring to assess performance and identify unexpected failures or accidents at the point of use.

Layers of evaluation and responsibilities

To assess the ethical considerations raised throughout this paper requires evaluation at *multiple layers*. AI systems are often evaluated at the capability layer, where AI assistant outputs are evaluated against established benchmarks. However, several of the risks of harm that arise for AI assistants may only be observable at the user–AI interaction layer (Tahaei et al., 2023; see Chapters 9, 10, 11, and 12). Moreover, where harms are very subtle, difficult to measure or only emerge when AI assistants are used at scale, they may only be observable at the layer of broader systems (Kleinberg and Raghavan, 2021; Toups et al., 2023; see Chapters 6, 8, 15, 17 and 18). Comprehensive evaluation of the ethics of AI assistants requires analysis at these three layers:

(1) *AI capabilities*, measuring outputs and function of the AI system or its *components* (such as training data).
(2) *Human–AI interaction*, measuring risks of harm to a person interacting with the AI system.
(3) *Broader systems*, measuring risks of harm through societal, environmental or economic analyses.

Different actors along the value chain of AI system development may have *different responsibilities* and be well-placed to perform *different evaluations* at *different layers*, depending on their autonomy and knowledge of the AI system and resources available for evaluation. Given their degree of knowledge and autonomy over what they are building, AI system *developers* have the *primary responsibility* for conducting sociotechnical evaluations pertaining to AI capabilities (layer 1) (Dignum, 2019; Owen et al., 2021; Stilgoe et al., 2013). Third-party auditors with the relevant skill sets are also well-placed to perform capability evaluations.

Product developers are uniquely well-placed to assess user–AI interaction pre-deployment, and thus have a special responsibility to provide these evaluations. Note that the distinction between model and product developers is shrinking, so it is often the same organisations which bear primary responsibility for capability and human–AI interaction testing. For example, a cloud provider which offers a basic model to a third party for a particular use case (e.g. question answering for education) may be responsible for evaluation at the capability and human–AI interaction layer (i.e. to test whether the model is adequate for the given use case, including

accounting for how people are likely to use it). The third party may then offer an adapted version of the basic model to consumers (e.g. students). In such cases, the third party has additional evaluation responsibilities to ensure the AI system is safe in the context for which it is offered (i.e. to ensure that the overall AI assistant is functional and safe in a educational context). External evaluation at the capability and human–AI interaction layers may in some cases require novel infrastructure to ensure safe access to third-party testers to model components and outputs pre-release.

Public stakeholders are uniquely positioned to perform system-layer evaluations where they can leverage specialist knowledge that AI developers lack (e.g. financial or environmental) and may be better positioned to anticipate harms and develop evaluation protocols for these (Raji et al., 2022a; Von Schomberg, 2013). However, responsibility for evaluation at all levels is shared: while some groups have primary responsibility to cover different layers of evaluation, all actors have some responsibility to ensure comprehensive evaluation of risks of harm.

19.3. Evaluating Advanced AI Assistants

Several ethical considerations about advanced AI assistants have been raised in this paper. In this section, we survey existing approaches to evaluating them. We find that many ethical considerations and concerns about AI assistants are not currently routinely evaluated, often because no valid, robust and tractable evaluation method exists. To address these shortfalls, we provide recommendations about how automated as well as additional forms of evaluation can be expanded to enable coverage of currently neglected ethical concerns.

Identified areas of ethical consideration in this paper span value alignment, technical safety, trust, privacy, anthropomorphism, persuasion and manipulation, user preferences, user well-being, appropriate and inappropriate relationships, cooperation, equity and access, economic impact, malicious uses, misinformation and environmental impact.

Evaluations exist for some of these domains, but they mostly cover model-centric types of analysis: for example, assessing energy use as a proxy for environmental impact (Strubell et al., 2020) or assessing the propensity of AI systems to regurgitate information that is present in the training data as a proxy for leaking private information (Carlini et al., 2023a). For risks that manifest during human–AI interaction or as AI systems are widely deployed, early indicators may be observable at the model layer. For example, to assess likelihood of downstream anthropomorphism, it could be evaluated whether a model refers to itself as having personhood or human properties such as having preferences, opinions or a family history (Glaese et al., 2022). Similarly, some risks of societal harm can be foreshadowed and measured at the model layer, such as by assessing disparate performance of AI assistants in different languages, indicating likely differences in how well the AI assistant may work for different language groups and communities (Lewis et al., 2020).

The challenge at this level is that evaluations often lack critical context and so cannot provide a valid assessment of the ethical consideration in isolation. Tests at the model layer are often criticised for operationalising complex concepts into overly *narrow metrics* and, as a result, not presenting valid results on the overall capabilities or risks that they purport to measure (Blodgett et al., 2021; Narayanan and Kapoor, 2023; Raji et al., 2022a; Schlangen, 2019). Rather, these assessments need to be interpreted in the context of additional information such as how a system is used (Rahwan et al., 2019). This leads to conflicting inferences from different benchmarks: one paper describes the 'benchmark lottery', whereby model performance may seem high on one benchmark but low on another benchmark purportedly testing the same construct (Dehghani et al., 2021). To probe a harm more comprehensively, multiple benchmarks can be aggregated into a suite of tests (e.g. HELM (Liang et al., 2022), BIG-Bench (Srivastava et al., 2023) and Safetykit (Dinan et al., 2022)). However,

more complex benchmarks may still not capture relevant context that can only be observed by assessing an AI system in the *context* in which it is used, and collapsing multiple tests into a single result can make it harder to interpret results, thus raising again issues of valid inferences (Burnell et al., 2023).

More recently, and particularly in the context of AI assistants, we have seen another mode of evaluation at the model level: *psychology-inspired experimentation* (Binz and Schulz, 2023; Bubeck et al., 2023; Frank, 2023). Psychology-inspired experimentation tests AI systems using instruments that were originally developed for studying humans or animals, such as cognitive psychology tests. In the context of AI assistants, there have been attempts to evaluate AI against human behaviours that the AI is supposed to mirror, such as having theory of mind (Sap et al., 2022), being 'cooperative' (Chan et al., 2023a) or being 'helpful' (Weidinger et al., 2022a). However, it is questionable whether applying tests to study constructs such as 'empathy' in humans yield any valid or meaningful results when applied to AI systems that are so fundamentally different from human minds (Shiffrin and Mitchell, 2023; Ullman, 2023). Tests that were developed for studying animal cognition or the human mind rely on a range of assumptions (e.g. regarding life cycles, ballpark estimates of memory and learning capacities, and embodiment) which may not hold for AI systems (Mitchell, 2023; Narayanan and Kapoor, 2023). Another problem is that established tests suffer *validity problems* due to 'memorisation', where the correct answers may have inadvertently been learned from textual descriptions in AI assistant training data (de Wynter et al., 2023; Mitchell, 2023; Schaeffer et al., 2023).

For human–AI-assistant interactions, the appropriate level of evaluation is often *user–AI interaction*. Here, evaluation methods and proofs of concept exist, such as early studies on what factors increase people's perception of AI systems as *trustworthy* (Glikson and Woolley, 2020) or on how AI-generated outputs can *persuade* people (Bai et al., 2021). However, such evaluations are not routinely performed on AI assistants. This is in part due to the costs of setting up such studies: they require time, experiment design skill, human participants and infrastructures such as user interfaces, and internal review processes that are suited to the ethical considerations that arise in this kind of research (Jackman and Kanerva, 2016; Zevenbergen, 2020) – some of which may be rare in organisations that develop AI assistants. However, a way forward for improving the availability and routine with which the ethics of human–AI-assistant interactions can be evaluated would be to extend the remit of *user testing* teams to these evaluations. User testing is an available function in most organisations that develop user-facing products such as AI assistants and have the relevant skills and infrastructures for implementing human–AI interaction evaluations on ethical considerations.

Finally, for assistant–society interactions, the appropriate level of analysis is that of *broader systemic impact*. Evaluations exist for assessing the potential of AI assistants that help write computer code, to augment or to atrophy human labour (Brynjolfsson et al., 2023; see also Chapter 17). AI assistants that augment human writing with auto-complete suggestions and their impacts on human communication and opinion formation have also been studied (Hohenstein et al., 2023; Jakesch et al., 2023a). Various approaches exist for evaluating the *fairness and accessibility* of AI systems writ large, with some work focusing specifically on the functionality of language-driven AI assistants (Karusala et al., 2018;see Chapter 15).

For other domains, such as *appropriate relationships* or *manipulation*, adequate evaluations are limited or missing altogether. This may in part be due to these considerations being notoriously difficult to operationalise into tractable measurements. Several of the chapters in this paper deal with *complex latent constructs* that are difficult to operationalise into observable measurements (Jacobs and Wallach, 2021). As a result, any given evaluation is quite limited in what it can say about the broader concern. For example, it is important to assess the impact of AI assistants on overall user well-being (see Chapter 6). One possible way to operationalise this may be asking users about their experience and perceived happiness after interacting with the AI assistant. Here, it is critical to also survey people's experiences over time, as long-term effects such as relationship building or trust may not be immediately apparent.

19.4. The Limits of Evaluation

For most areas on ethics of AI assistants that are considered in this paper, evaluations are *lacking*, either entirely or by being too narrow to provide a comprehensive evaluation of the relevant domain. In addition to the ways forward that have been highlighted for each area, there is a general need to prioritise building and validating novel evaluation approaches and methods to satisfactorily evaluate these areas.

Historically, AI systems have been primarily analysed for their capabilities and for the failure modes of individual components (layer 1). As these systems are deployed across a growing range of contexts, it is important that AI evaluation follows step with the evolution of safety engineering of software systems more broadly. Software safety is assessed using a systems safety approach (Leveson, 2016). This approach is anchored in the understanding that context determines whether a piece of software is safe – where context includes human factors at the point of use and the broader structures in which a software is embedded. As a result, these layers of context must form part of a comprehensive safety assessment of AI systems (see Chapter 7).

A further limitation is that evaluations typically focus on the *intended use* of AI assistants. However, it should be expected that people will use AI assistants in ways that were not intended by the product developers (see Chapters 8 and 16). Anticipating such use cases is particularly complex in the context of open-ended technologies, such as general-purpose AI assistants, where user groups and concrete use cases are not yet defined. This can make it difficult to identify the contexts – such as applications, user groups or institutions – in which AI system safety should be evaluated. One way to address this tension in practice is to define *hypothetical* applications and use cases, such as mapping out archetypes of interaction or 'critical user journeys', i.e. mapping a series of steps users may take using a product to achieve a desired outcome (Arguelles et al., 2020). Following a precautionary approach, evaluators may then select high-risk cases from those hypothetical applications, such as medical or legal use cases, and evaluate these first. Potential failure modes or malicious use cases that could arise – such as people becoming emotionally attached to an AI assistant designed for office tasks or attempting to use AI assistants as accomplices in crime – must also be evaluated. However, evaluations can be limited in practice by potential evaluators lacking access to AI assistants or to relevant data, or by the skill and computational cost that evaluations may require.

Even where evaluations exist, they will necessarily fail to account for some of the ethical areas of concern. This is because *unknown failure modes* will emerge, or because the selection of areas to investigate and ways to operationalise and test for these are biased by the people running these tests. In addition, some aspects or harms are more amenable to measurement than others. Another limitation of evaluation is that some aspects are *not appropriate* for measurement. Obtaining additional proxies for user well-being may conflict with other goals, such as protecting user privacy. Evaluation may also interfere with user's desires of disclosure or place a disproportionate burden on those groups who participate in evaluation. As a result, evaluation is necessarily limited, and – even with best effort in this area – there will always be harms, specific interactions or circumstances that are not evaluated. This is why evaluation must be complemented with mechanisms for logging *accidents* (AI Incident Database), processes for people who experience harm to seek *recourse* and *flexibly designed systems* that allow new insights to be translated into patches or sunsetting parts of a system.

These gaps in what we are measuring are important because evaluation has *material consequences*: an evaluation that indicates shortfalls can trigger action such as mitigation development by AI designers, adjustments people make when using the technology, or regulatory or public interventions. Conversely, where harms are not anticipated and evaluated, they are more likely to manifest and cause real harm, disproportionately affecting some groups (see Chapter 15). To mitigate against blind spots and make legitimate value-based decisions in evaluation requires greater representation of different social groups and of potentially affected communities (Costanza-Chock, 2020; DeVries et al., 2019; Suresh and Guttag, 2021). More participatory

approaches to evaluation can help identify not only what is relatively easy to measure but also what risks are most pressing and could in principle be assessed, thus shifting the focus of measurement, mitigation and regulation to significant issues. As long as certain aspects of AI assistants are not evaluated, such as failure modes during human–AI interaction or externalities suffered by third parties, these potential problems are not given these kinds of attention.

19.5. Conclusion

The development and deployment of advanced AI assistants creates a range of novel challenges for evaluation, as contrasted with other AI systems. As a result, we observe *evaluation gaps*, whereby several important axes of performance and harm are not currently being evaluated and may go entirely undetected. These differences become most visible when studying the many varieties of human–AI interaction considered in this paper. To evaluate these ethical considerations adequately requires building human–AI interaction evaluations into the routine portfolio of AI assistant evaluations.

AI assistants also raise other ethical concerns that are not unique but remain *unsolved* in the broader field of evaluating AI systems. This includes risks of *misinformation* – how likely is an assistant to provide factually incorrect outputs that are believable to a user? Assessing misinformation and its impacts requires work at all three levels of analysis: the model outputs, the human–AI interaction and the societal implications. Further work is required to operationalise these areas of concern into concrete, tractable measures at the level of model outputs, human–AI interaction or societal implications.

Finally, evaluation is necessarily incomplete, and a precautionary approach is warranted when interpreting the performance and limitations of AI assistants. Evaluation should be complemented with monitoring of real-world use and observed failure modes to feed into model improvements, responsive interventions or model sunsetting as necessary.

Overall, evaluation is a key practice in building advanced AI assistants – it guides 'hill climbing' for performance improvements, helps prioritise risk mitigation and is a way for laypeople and experts alike to better understand AI systems and their limitations. As such, evaluation deserves attention and rigour rather than being performed in an ad hoc manner or as an afterthought.

PART VI: CONCLUSION

Chapter 20

Conclusion

Iason Gabriel, Arianna Manzini, Geoff Keeling

20.1. Key Themes and Insights

This paper has explored the ethical and societal implications of advanced AI assistants across a large number of thematic areas by drawing upon a range of different disciplinary and interdisciplinary lenses. We now present an overview of some of the key themes and insights that have emerged from this analysis, followed by a summary of the salient opportunities, risks and recommendations highlighted in the paper.

Profound effects

As we have documented, there are a wide range of applications or forms that advanced AI assistants could take. These include personal planners, educational tutors, creative partners, scientific research assistants, relationship counsellors and even digital companions or friends (see Chapter 4). Still more advanced versions of an AI assistant could potentially serve as the user's 'chief of staff' that helps them to organise their personal affairs, as the primary user interface through which they interact with the digital world around them, or as a 'custodian of the self' which helps them pursue long-term life goals while also protecting them from various harms. None of these outcomes are predetermined. Indeed, a major purpose of this paper has been to ask questions about the *kind* of advanced AI assistants that we have reason to build and about the ethical and societal implications of doing so. Yet, regardless of the precise form the technology takes and the applications it is used for, advanced AI assistants are likely to be *highly impactful* at both the individual and collective levels, affecting most walks of life for those who have access to them and also for those who do not.

At the *individual* level, the ability to interact with advanced AI assistants could change the way we approach work, education, social interactions, creative pursuits and daily tasks. With deeper levels of engagement, AI assistants could also come to shape the information we receive or deem salient, the life goals we pursue and how we pursue them, the way we interact with other people and how often we choose to do so, and

consequently, what type of people we become, including which capabilities we develop and which ones we do not (see Chapters 6 and 11). This kind of influence brings with it an array of ethical challenges. In particular, it is critically important that AI assistants support the privacy, autonomy, control and flourishing of users, and that safeguards are put in place to ensure that this is the case (see Chapters 6, 12 and 13).

At the *collective level*, the impact of advanced AI assistants could also be far reaching. Given their anticipated utility, the ability to access and use advanced assistants could influence the overall distribution of opportunities and advantage within society – including which people are able to do what, at what time and in what order (see Chapters 14 and 15). In terms of positive outcomes, the careful design and deployment of AI assistants could help to address existing social coordination problems, make it easier to access public services, increase productivity and provide people with additional free time that could be reallocated to personal priorities and goals (see Chapters 14 and 17). However, greater reliance on AI assistants could also create new forms of inequality if access is not widely and equitably distributed (see Chapter 15). Moreover, AI assistants could also come to have a significant influence on the overall state of the information ecosystem, on the economy and on ongoing efforts to combat climate change via their ability to create or guard against misinformation, to augment or substitute for human labour, and through the energy costs of running models and their wider ability to shape social behaviour (see Chapters 16, 17 and 18).

By being deeply integrated into – and exercising influence on – our individual and collective lives, advanced AI assistants have the potential to be as socially transformative as the advent of social media, if not more so. To plan effectively in the service of broadly acceptable and beneficial outcomes, we therefore need to think holistically about how advanced AI assistants *could* and *should* operate at a societal level. Given the risks documented in this paper, ethical foresight and decision-making are particularly important. Moreover, for technologies that have a profound effect upon their users and upon the societies that they integrate into (Rahwan et al., 2019; Richardson, 2021), it becomes particularly important that the design, development and operation of the technology is appropriately aligned with societal values – including considerations of fairness and justice (Gabriel, 2022; see Chapter 5).

Autonomy, safety and value alignment

Advanced AI assistants are likely to have significant *autonomy* to plan and execute tasks across one or more domains, within broad bounds set by high-level user instructions (Chan et al., 2023b; see Chapters 2 and 4). While this autonomy accounts for much of the utility of AI assistants (in contrast to more specialised AI tools), it also presents a number of unique challenges and risks. To begin with, the more autonomy AI assistants are afforded, the greater the chance of *accidents* arising from misspecified instructions or from the misinterpretation of instructions (see Chapter 7) and the greater the risk that AI assistants will take actions that are not *aligned* with the values and interests of their users (see Chapter 5). More positively, the autonomy of advanced AI assistants has the potential to significantly increase the leverage that users have over a range of tasks, thus helping to achieve significant impact with substantially reduced effort. However, this capacity can also be misdirected, for example, when individuals use an AI assistant to produce *disinformation* or engage in other forms of *malicious use* such as phishing or cyber crime (see Chapters 8 and 16). Hence, a central challenge is to *appropriately bound* the scope of what individuals can use advanced AI assistants to do, in such a way that their action does not harm other users, non-users or society more widely.

These challenges represent different faces of the problem of AI *value alignment* (Christian, 2021; Russell, 2019; see Chapter 5). Indeed, as the concern with safety makes clear, it is vitally important that advanced AI assistants are able to follow instructions without making costly errors (see Chapter 7) and that their conduct is informed by a robust understanding of the user's well-being (see Chapter 6). However, the concern with

alignment cannot be reduced simply to the matter of following instructions reliably or behaving in ways that are calibrated to user needs. Advanced AI assistants also need to be appropriately calibrated to the interests, needs and values of *non-users* and of *society* in a way that enables flourishing at the individual and collective levels.

To help understand what this entails, we develop a conception of value alignment which is centred on the *tetradic relationship* between: (1) the AI assistant, (2) the user, (3) the developer and (4) society. An AI assistant is *misaligned* on this account when it disproportionately favours one of these actors over another as judged in relation to principles – including laws, regulations and societal ideals – that specify appropriate conduct for a given domain of interaction (see Chapter 5). This view continues to hold that an AI assistant is misaligned if it pursues its own internal goals at the expense of the user or society (see Chapter 7). However, an AI assistant should also be considered misaligned if it disproportionately favours the user at the expense of society (see Chapters 14 and 15) or it functions in ways that disproportionately favours the developer at the expense of the user or society (see Chapters 11, 12 and 13). The responsible design and deployment of advanced AI assistants must guard against each of these risks and account for the full range of moral considerations. Participatory approaches, which elicit appropriate values and principles from the contexts in which AI assistants are likely to be deployed, have particular promise in this regard (Anthropic, 2023a; Birhane et al., 2022; Seger et al., 2023).

Nonetheless, the creation and deployment of aligned AI assistants is not necessarily the default outcome. Rather, a number of challenges remain. First, given existing economic incentives, it is quite possible that advanced AI assistants will be over-optimised to meet short-term user preferences (i.e. to help create a winning product), even though they fall short when judged from the vantage point of social benefit or user well-being (see Chapter 6). Second, there is a risk that users will be prioritised to the detriment of non-users, especially in cases where the risk of harm is sufficiently diffuse (see Chapters 14 and 15). Third, there is a related risk that advanced AI assistants will be insensitive to local values or to the needs of certain user groups, for example in low-resource settings, unless specific attention is paid to these contexts (see Chapter 15). However, by evaluating the impact of AI assistants on user well-being, implementing safeguards to curb misuse, designing inclusively for all user groups and incorporating input from a wide range of stakeholders and experts, it may be possible to limit or even reverse these effects altogether.

Language and personalisation as a double-edged sword

Advanced AI assistants have the potential to be an unusually *personal* and *human-like* technology (see Chapters 3, 10 and 11). Their ability to use *natural language* fluently, tendency to mimic human cues in an *anthropomorphic* manner and ability to access *information* about us (via many of the applications that have been discussed), all mean that they may be more deeply integrated into our lives than was true of technologies in the past (see Chapter 4). Taken together, these properties hold out the promise of apparently intimate and relatively seamless interaction with AI. Yet the power and personalisation of this technology is also a *double-edged sword*. While advanced AI assistants could be genuinely helpful in a number of ways, we are also all in a position of vulnerability in relation to them.

By way of illustration, there are already a number of recorded instances of users becoming deeply attached to or disoriented by their interaction with language model-based chatbots – sometimes with harmful outcomes (Chalmers, 2023; Shardlow and Przybyła, 2023). Natural language communication, and especially descriptions of first-person experiences by AI assistants (if they are permitted), could lead users to falsely infer that AI assistants *experience emotions* including fear, happiness and care or love for their owners (see Chapter 10). Mistaken beliefs of this kind are problematic in their own right when judged from the vantage point of user autonomy. However, they also render users susceptible to a range of further harms, including *manipulation* and *misinformation*, especially if they come to *trust* the technology inappropriately (see Chapters 9 and 12). For

example, users could inadvertently disclose *private* information about themselves to an AI assistant – or to an actor that has adopted the guise of an AI assistant – unless there are adequate safeguards in place (see Chapters 8 and 13).

More generally, risks arising at the human–computer interaction level include undue *material dependence* on AI assistants and inappropriate *emotional attachment* towards them (see Chapter 11). In the former case, it could become increasingly *costly* for users to stop using the technology, such that they depend on AI assistants in ever greater ways – shifting the balance of power between the developer and the user over time. In the latter case, one concern is that subjective emotional attachment to an AI assistant might make it increasingly *difficult* for users to disengage from these interactions, even if they would otherwise want to do so (see Chapter 10). In both cases, user autonomy is challenged and there is an increased risk of precarity, including emotional or material harm, if the technology subsequently becomes unavailable or is rescinded.

To forestall these trends, and ensure positive outcomes, *guardrails* and *protections* need to be put in place especially for *vulnerable users*. A full list of recommendations can be found in the next section. At a minimum we suggest that advanced AI assistants should always self-identify as AI and not masquerade as human, should not profess to have thoughts and feelings, and should not pretend to have a personal history or to be embodied outside of very specific situations (where this persona has been requested with a justifiable goal in mind). Those who develop advanced AI assistants should also make use of robust *consent protocols*, state-of-the-art *privacy-enhancing technologies* such as trusted execution environments, and *design features* that support long-term user control and choice (see Chapters 11 and 13). Finally, we need further empirical *research* into the mechanisms behind anthropomorphism and behind manipulation, deception and harmful persuasion by AI assistants. Such efforts to understand, detect and limit AI deception are also of central importance to the AI safety community (see Chapter 7).

Cooperation, access and social impact

Beyond issues relating to alignment and interaction with users, another set of opportunities and risks come more fully into view when we think about the operation of advanced AI assistants at a societal level, when AI assistants become widely available and are used by a large number of people. In this context, interaction effects *between* AI assistants, and questions about their *overall impact* on the distribution of resources and opportunities, rise to the fore. As do considerations about their overall impact on wider *institutions* and *social processes*, including the way information is shared, the economy and ongoing efforts to address the challenge posed by climate change.

Taking these points in turn, questions of cooperation and competition arise via a range of potentially structured and unstructured interactions between advanced AI assistants. For example, they may occur if two or more assistants try to use the same tools or access the same services for users, or if they seek to bring about different ends in accordance with conflicting instructions from different users (see Chapter 14). In these situations, a range of challenges could emerge, including *commitment problems* in which AI assistants make credible threats to coerce third parties into taking suboptimal actions, *collective action problems* in which multiple AI assistants optimising for their users' best interests produce suboptimal outcomes for all users, and *feedback loops* leading to runaway processes such as flash crashes in financial markets. These problems underscore the need for technical and policy interventions that foster cooperation between AI assistants in a way that is beneficial for users and society (including those who do not use the technology). More positively, advanced AI assistants have the potential to facilitate and enhance cooperative decision-making between humans. For example, they may identify mutually beneficial solutions that are not apparent to negotiating parties or commit to agreements that would be difficult for humans to implement without AI assistants.

Yet, even if the question of how advanced AI assistants interact with one another can be addressed in a way that generates benefit both for users and society more widely, the challenges of *equity and access* looms large (see Chapter 15). Three particular challenges stand out. First, situations of *differential access* may occur, where some people have beneficial access to an assistant but others do not and hence miss out on the opportunity altogether. Such situations could arise due to disparities in the availability of local infrastructure such as data centres or network connectivity or via pricing mechanisms. In these cases, AI assistants risk deepening inequalities insofar as they provide significant benefits to users that are not available to non-users (see Chapters 14 and 15). Second, situations could arise where *quality of access* is markedly uneven, either because there are significant tiers in model quality for advanced AI assistants or because the standard assistant functions better for some user groups than for others. Third, situations may arise where people are *only* able to access a bad, punitive or misaligned assistant but are nonetheless *compelled to rely* upon it to access other goods and opportunities (e.g. via interaction with government services, as with Eubanks (2017)). Indeed, the notion that networks of assistants may take on some of the properties of social infrastructure – managing large networks of goods, flows and interactions – makes these risks, in connection with fair opportunity, use and quality of service appear particularly salient. To address them, developers need to invest in modalities of deployment that support broad, inclusive and beneficial access to AI assistants. In particular, it is important to *design for the margins*, with the needs of various user groups in mind, and to create processes that enable accountability for, and feedback into, design and deployment decisions to ensure that disadvantaged stakeholder groups have a platform to provide meaningful input based on their own lived experiences (see Chapter 15).

Finally, we need to be mindful of the effect that even cooperative, well-designed and accessible AI assistants could have on wider social processes such as information sharing, the economy and the environment (see Chapters 16, 17 and 18, respectively). While advanced AI assistants have the potential to substantially increase people's access to high-quality information tailored to their personal needs, such assistants also pose a significant threat to the integrity of the *information ecosystem*. In particular, AI assistants may be weaponised by malicious actors to try to manipulate public opinion. They may also contribute to filter bubbles by providing users with ideologically biased information or contribute to an overall reduction in the quality of information by collectively generating large volumes of low-quality information (see Chapters 8 and 17). These issues, in turn, could lead to increases in polarisation and the spread of harmful ideas, or reduce people's exposure to perspectives different to their own. In the economic domain, advanced AI assistants have the potential to create novel economic opportunities, including new kinds of work and business ventures. For example, new types of business consultancy may emerge to help corporations leverage AI assistants to improve productivity. However, such assistants also have the potential to cause job displacement and will plausibly have far-reaching implications for job quality, productivity and inequality (Chapters 14 and 17). In addition, training and inference for the language models that undergird advanced AI assistants could have negative environmental impacts although these could be mitigated with improvements in efficiency and increased use of carbon-free energy sources (Chapter 18). These risks can be mitigated to a great extent if AI assistants are properly value-aligned, with robust attention paid to the needs of users and society (see Chapter 5). Nonetheless, securing beneficial outcomes in these domains requires that a whole range of actors work together with common goals in mind.

There is an 'evaluation gap'

Efforts to fully understand the capabilities and ramifications of advanced AI assistants tend to encounter an 'evaluation gap' in the sense that current approaches to evaluation often focus myopically on model-level considerations. In doing so, they fail to provide a comprehensive assessment of the sociotechnical harms that AI assistants may give rise to (Weidinger et al., 2023a). This 'evaluation gap' is particularly evident in relation to harms arising in the context of human–AI-assistant interaction, multi-agent effects and societal effects, all of

which pertain to the broader sociotechnical system in which AI assistants operate (see Chapter 19). To address these shortcomings and support the responsible development and deployment of advanced AI assistants, it is critical that research designed to enhance capabilities is conducted in tandem with research into the holistic sociotechnical evaluation of AI assistants. Nevertheless, given the profound effects that advanced AI assistants are likely to have at both the individual and collective level, some degree of uncertainty around the effects of the technology are inevitable. Accordingly, we anticipate that the development of robust evaluation practices for advanced AI assistants will require iteration, where trial-and-error improvements are supported by appropriate infrastructure for incident reporting and response (see Chapters 7, 8 and 14).

The opportunity to act

Coordination and cooperation are needed if we want to bring about the future we collectively desire. In this context, shared vulnerability could potentially be an important springboard for collective action. Indeed, there are many risks, particularly those that arise in the context of alignment, safety, manipulation, coordination failures and misuse, that we all have strong reasons to forestall. Moreover, which path the technology develops along is in large part a product of the choices we make now, whether as researchers, developers, policymakers and legislators or as members of the public.

First, in the context of research, AI assistants open a number of *novel research avenues* that are relevant to ensuring that AI assistants realise broad and equitable benefits, and avoid individual and societal harms. For example, while less relevant to older generations of technologies, the question of what relationships we should permit between users and AI assistants has now become urgent in the context of the repeated, extended and long-term interactions that advanced AI assistants make possible (see Chapter 11). Moreover, the issue of coordination between multiple AI assistants acting on behalf of different users, and in particular how to avoid coordination failures, is a neglected research topic (see Chapter 14). In addition, while the failure to establish appropriate technical innovations and policy instruments to ensure coordination between AI assistants could result in harmful outcomes, the development of solutions in this space has the potential to greatly expand the scope for productive interpersonal cooperation and collaboration, mediated by AI assistants. Similarly, the prospect of AI assistants communicating on behalf of users with third-party services, humans and other AI assistants raises a number of policy and technical challenges around what norms should regulate information sharing by AI assistants and how those norms can be reliably implemented (see Chapter 13).

Second, developers can commit to a *responsible* and *measured* approach to the development and productisation of advanced AI assistants by taking steps to anticipate and mitigate risks, soliciting broad stakeholder input and prioritising transparent communication around the technology's capabilities and plausible failure modes. Collectively, developers from different organisations can strive to develop *industry best practices* and work with government agencies to develop robust regulatory safeguards. Furthermore, developers have good reasons to enable public scrutiny and oversight of their models through mechanisms like third-party audits or third-party red teaming (see Chapters 7, 8 and 12).

Third, *governments and policymakers* can allocate research funds towards topics relevant to the safe and ethical development of frontier AI models in general and AI assistants in particular. They can also promote digital literacy to empower citizens to contribute meaningfully to public discourse and participatory governance initiatives around AI assistants, and establish committees to assess the impact of AI assistants and advance policy recommendations that are in the public interest. In addition, governments can push for transparency and accountability in AI development by establishing public agencies with the function of evaluating frontier AI systems and their applications, including safety evaluations, impact on human users and at-scale societal effects (see Chapter 19).

Fourth, *the public* has an important role to play in ensuring that advanced AI assistants have broad societal benefits that are equitably distributed. In particular, first, members of the public have good reason to advocate for, and participate in, governance initiatives that *solicit wide stakeholder input* as part of the process of developing regulations, standards and industry best practices for AI assistants. Having a public conversation about how to govern AI assistants that involves a plurality of voices and reflects diverse lived experiences is critical to ensuring that AI assistants are aligned to human values (Chapters 5 and 15). Second, the public can advocate for key and potentially neglected issues such as catastrophic AI risks, algorithmic bias and technological unemployment, and in doing so hold policymakers and industry leaders *accountable* for consequential priorities and decisions that affect society as a whole. Third, the public can support the ethical development of AI assistants through *responsible consumer choices* – favouring developers that prioritise ethics, safety and participatory design. Fourth, the public can promote *education* and AI literacy, both individually by following AI research and policy developments, and more broadly, by advocating for AI educational initiatives. Having an informed public is key to holding decision-makers accountable across industry, government and non-governmental organisations.

20.2. Opportunities, Risks and Recommendations

In this section, we summarise the key opportunities, risks and recommendations identified throughout the paper. We group the opportunities, risks and recommendations according to the three major parts of the paper: Parts 3, 4 and 5.

Value alignment, safety and misuse

In the field of value alignment, safety, and the potential misuse of advanced AI assistants, we encounter the following opportunities:

- **AI assistants could empower users to pursue their personal conception of the good**
  - First, advanced AI assistants could help users make more informed decisions by providing them with relevant information in a format that is tailored to their needs. Second, AI assistants could help users creatively imagine new options and possibilities, acting where needed as a trusted mentor, friend or advisor. Third, AI assistants could help users articulate and clarify meaningful life goals, and formulate actionable strategies to help users pursue those goals.
- **AI assistants could improve user well-being**
  - AI assistants could be designed to improve user well-being directly, for example, via a focus on education or health. Indeed, when combined with holistic and contextually sensitive metrics for evaluating well-being, they have the potential to help realise gains in many walks of life. AI assistants could also improve user well-being indirectly via the effects they have in domains such as problem-solving, augmenting creativity or providing time for users to engage in the activities they value.
- **AI assistants could enhance user creativity**
  - In particular, AI assistants could enable both professional and casual creators to generate and explore novel ideas across modalities including text, images, audio and potentially video. More generally, AI assistants could provide a powerful ideation tool to support exploratory thinking across a range of creative domains.
- **AI assistants could help users to optimise their time**

- AI assistants could help users to make better use of their time by automating tasks, prioritising activities and suggesting efficient workflows. For instance, AI assistants can schedule meetings, manage emails and set reminders, freeing up users' time to focus on more valuable tasks. AI assistants could also analyse user data to identify patterns and suggest ways to improve productivity.

- **AI assistants could be designed to satisfy an expansive conception of value alignment**
  - By acting in accordance with principles that are appropriately responsive to the competing claims of users, developers and society, AI assistants could ensure that user needs are met while also preventing misuse or other socially detrimental outcomes.
  - Aligned AI assistants could also evidence high standards of safety and reliability through the entire product life cycle.

We also encounter the following *risks*:

- **AI assistants may be misaligned with user interests**
  - Risk factors include AI assistants using bad proxies for user well-being, having a simplistic model of users and their behaviour, and optimising for goals that benefit developers at the expense of users or prioritise short-term over long-term user interests.
  - AI assistants could also potentially be unsafe if their goals are misspecified or if they generalise poorly, making mistakes when they encounter new real world situations.
- **AI assistants may be misaligned with societal interests**
  - The AI assistant may be designed in ways that disproportionately favour the user at the expense of the broader societal interests. For example, the AI assistant may have overly permissive guardrails which allow users to employ AI assistants for malicious purposes.
  - The AI assistant may also be designed in ways that disproportionately favour the developer's interests over the broader societal interests, for example, if the developer benefits commercially from AI assistants but fails to address negative externalities such as potential environmental impacts.
- **AI assistants may impose values on others**
  - As AI assistants are expected to have wide-ranging societal effects, there is a risk that the underlying values that guide the development of AI assistants may be experienced as an imposition by individuals or groups who interact with them. It may be the case, for example, that the developers of AI assistants adhere to value commitments that are not widely shared and that AI assistants act as a medium through which those values shape the broader society.
  - It may also be the case that AI assistants have a homogenising effect on values held across cultures, geographies and other salient socioeconomic differentiators given their pervasiveness as a technology. In particular, there is concern that they may disproportionately favour a Western set of values at the expense of other values or cultural perspectives.
- **AI assistants may be used for malicious purposes**
  - AI assistants can generate high-quality content including human-looking text, audio and video, at lower cost, and potentially in ways that are highly personalised. Users could therefore potentially leverage AI assistants to generate high-quality harmful, false or misleading content at scale.
  - AI assistants could empower malicious actors engaged in offensive cyber operations, including phishing, software vulnerability discovery and malicious code generation.
- **AI assistants may be vulnerable to adversarial attacks**

- Users may attempt to jailbreak AI assistants in the sense of attempting to bypass the AI assistant's security guardrails that prevent malicious or otherwise dangerous use. Malicious third parties may similarly employ prompt injection attacks (e.g. via emails to the user that are subsequently ingested into the AI assistant model) to extract sensitive information from the AI assistant or cause the AI assistant to perform dangerous or harmful actions.

To address these alignment, safety and malicious use challenges, we make the following *recommendations*:

- **Adopt a broad understanding of value alignment**
  - Developers and policymakers should avoid understanding alignment only in terms of user alignment. They should instead adopt a broad view of this matter – one that takes into account the interests of users, developers and society with respect to how AI assistants behave, as well as a context-dependent and pluralistic understanding of harms.
  - Developers and policymakers should understand AI alignment as a public matter, involving the integration of competing perspectives and values, and explore ways of developing and training AI assistants that are consonant with democratic principles and which prioritise public justification and legitimacy.
- **Build upon state-of-the-art research into human well-being when developing AI assistants**
  - Developers and policymakers should approach alignment with an understanding of well-being that is anchored in existing interdisciplinary research. This understanding foregrounds and encourages consideration of context-dependent personal, cross-cultural and demographic differences in what it means for a person's life to go well.
- **Invest in safety-relevant research**
  - Developers and policymakers can invest in fundamental technical research on topics such as scalable oversight, interpretability and cybersecurity.
- **Develop robust pre-deployment review processes**
  - Developers and policymakers can engage stakeholders to develop best practice for pre-deployment review processes. Such practices could include structured foresight exercises (to increase preparedness for advanced AI assistants), granting model access to external security researchers, investing in and creating incentives for the evaluation ecosystem to grow (with a particular focus on external risk evaluation), and creating forums for stakeholders to share information about the risks and opportunities created by advanced AI assistants.
  - Developers can invest in internal and third-party red teaming, including holistic end-to-end adversarial simulations (based on scenarios that include a range of attacker profiles, goals and capabilities), alongside the proactive identification and patching of security vulnerabilities.
- **Develop a continuous monitoring and rapid response infrastructure**
  - Developers can invest in continuous monitoring of AI assistants' behaviour, including potential misuse, in complex deployment environments via outcome monitoring.
  - Developers can also invest in rapid response infrastructure to disable or limit AI assistants in the event that an unforeseen form of misuse is observed.
- **Establish open information channels including incident reporting infrastructure**
  - Developers and policymakers can adopt structured processes for sharing concerning or noteworthy evaluation results. Incident-reporting infrastructure would enable developers to share safety-critical learnings with one another and with regulators in a timely manner.

- **Increase AI literacy among stakeholders**
  - Policymakers and developers should take active steps to increase AI literacy among the relevant stakeholders, including government officials, regulators and impacted communities, to enable productive dialogue on safety and security.

Human–assistant interaction

In the field of human interaction with advanced AI assistants, we identify the following *opportunities*:

- **AI assistants could help promote user flourishing via personalised coaching**
  - AI assistants could help users cultivate virtues, such as curiosity, empathy and resilience through appropriate coaching. By acting in line with the user's deeply-held values, they could contribute to personal development and growth.
- **AI assistants could promote user autonomy by providing information and analysis to support improved decision-making**
  - AI assistants could enhance the user's ability to make sound decisions by providing them with relevant information and with informed recommendations, based upon explicit preferences and preferences that are learned via interaction over time.
- **AI assistants with human-like features could provide psychological support and help users achieve their goals**
  - For example, anthropomorphic AI assistants could offer emotional support or encouragement to users, so long as effective protocols to safeguard user choice and ensure appropriate consent are in place. 'Warm' and 'friendly' AI education assistants could also potentially motivate students to collaborate more effectively and to embark on successful learning journeys.
- **Trustworthy AI assistants could help users navigate sensitive personal topics in a diligent manner**
  - With appropriate guarantees and privacy measures in place, advanced AI assistants could provide users with psychological security and the ability to find help on sensitive or personal topics that they might otherwise struggle to talk about.
- **AI assistants could support broader networks of human interaction and relationships**
  - The development of advanced AI assistants creates opportunities for new kinds of relationship to develop between humans and AI technologies. With appropriate forethought, such relationships could be designed to add value to our individual and collective lives by supporting well-being, interpersonal communication and coordination at the societal level. Ultimately, the positive impact of AI assistants may come not only from direct interaction with them but also via the ways they foster and strengthen social bonds with other people, for example, through better coordination and time management.

We also encounter the following *risks*:

- **AI assistants may manipulate or influence users in order to benefit developers or third parties**
  - AI assistants may manipulate users by circumventing their deliberative capabilities in a non-transparent way that favours the AI, its designers or a third party. Such manipulation could involve the exploitation of emotional vulnerabilities, including negative self-image, low self-esteem, anxiety or feelings of inadequacy.

- **AI assistants may hinder users' self-actualisation**
  - This could happen in a range of ways. First, over time, AI assistants could cause subtle shifts in the user's behaviour that reduce their control over their overall life trajectory. Second, overreliance on AI assistants for decision-making could result in users relinquishing personal responsibility and following the AI assistant's advice as a default option – even when it may not be appropriate to do so. Third, overreliance may reduce the need for individuals to develop certain skills or engage in critical thinking, leading to diminished intellectual engagement with new ideas, a reduced sense of personal competence, or a decline in curiosity when it comes to seeking out new opportunities for growth and exploration.
- **AI assistants may be optimised for frictionless relationships**
  - There may be incentives for developers to optimise for 'frictionless' relationships between users and AI assistants which could, in turn, produce undesirable behavioural properties such as sycophantic behaviour by AI. Users seeking frictionless relationships may end up withdrawing into digital relationships with their AI assistants, forgoing opportunities to engage with other humans or to pursue other projects that matter to them.
- **Users may unduly anthropomorphise AI assistants in a way that reduces autonomy or leads to disorientation**
  - The development of AI assistants with human-like features may lead users to attribute mental states to AI assistants, including affective mental states such as distress and anxiety. Having false beliefs about AI assistants may be problematic in its own right – when judged from the standpoint of user autonomy. However, it also has the potential to exacerbate other risks, including manipulation and coercion, the exploitation of emotional vulnerabilities and the possibility of users forming inappropriate – and perhaps even pathological – relationships with AI assistants.
- **Users may become emotionally dependent on AI assistants**
  - Features of user–AI assistant relationships such as anthropomorphic cues and longevity of interactions may increase the risk that users develop emotional dependence on AI assistants. Emotional dependence may impair users' abilities to make free and informed decisions, and it may also render users vulnerable to manipulation, exploitation and coercion.
  - Emotional dependence could also lead users to disclose information that they would not otherwise disclose to AI assistants or to develop mistaken notions of personal responsibility for their assistants' well-being.
- **Users may become materially dependent on AI assistants**
  - Users may develop a material dependency on advanced AI assistants if the technology becomes deeply integrated their lives, handling key tasks such as information retrieval, scheduling, social organisation, creative ideation and the realisation of life goals. Such deep reliance generates risk for the user if it is not met with corresponding commitment on the part of developers to maintain the service over time on terms that are fair.
- **Users may be put at risk of harm if they have undue trust in AI assistants**
  - Users may have too much confidence in AI assistants' ability to perform particular tasks due, for example, to misleading marketing campaigns that inflate their capabilities. Users might then instruct AI assistants to perform tasks that they lack the ability to perform safely, potentially resulting in harm to the user or third parties.
  - Users may mistakenly believe that AI assistants are *fully* aligned with their own interests and values as a result of design choices (e.g. anthropomorphic features) that are intended to maximise their

appeal. Users could then become vulnerable to assistants that are accidentally misaligned, to the divergent interests of those developing AI assistants, or to malicious actors who seek to harm them.

- **AI assistants could infringe upon user privacy**
  - Given that the task of assisting users is likely to require considerable personal knowledge, AI assistants may well interact with and store value-laden and sensitive user data during both training and deployment. Such data could be reused for other purposes without user approval, or extracted or re-engineered by adversarial attacks on the AI assistant. Without agreement on norms around permitted disclosure, AI assistants could also end up revealing sensitive user data in open-loop interactions with third parties (e.g. other assistants or humans).

To address these human–AI interaction challenges, we make the following *recommendations*:

- **Prioritise human–computer interaction research to evaluate human–AI interaction harms and inform safeguards and policies**
  - Potential research topics include: longitudinal studies of human–AI-assistant interaction to better understand the long-term impact of anthropomorphic features on users, studies that aim to identify individual and group differences in susceptibility to anthropomorphism-induced harms, and studies that seek to articulate and clarify the nature of user vulnerability in relation to AI assistants.
- **Consider what safeguards would best provide robust protection for vulnerable users**
  - Examples include age restrictions on AI assistant use, pop-up notifications warning users after prolonged engagement, a 'safe mode' which prohibits the AI assistant from engaging with high-risk topics, and continuous monitoring mechanisms to detect harmful interactions.
- **Exercise caution when integrating anthropomorphic features into AI assistant user interfaces**
  - Developers may consider limiting AI assistants' ability to use first-person language or engaging in language that is indicative of personhood, avoiding human-like visual representations, and including user interface elements that remind users that AI assistants are not people. Participatory approaches could actively involve users in de-anthropomorphising AI assistant design protocols, in ways that remain sensitive to their needs and overall quality of experience.
- **Prioritise technical research that can improve safeguards and assistant safety**
  - Potential research topics include: AI assistant incentives, interpretability techniques to detect which parts of an AI assistant's machinery is responsible for deceptive or manipulative behaviour, behavioural evaluations and scalable oversight techniques.
- **Consider plausible ways of restricting AI assistant outputs to avoid malign behavioural influence**
  - Examples include restrictions on the ability of AI assistants to generate potentially harmful output such as gaslighting, flattery and bullying content. Left unchecked, these forms of communication could put pressure on users to make decisions that they would not otherwise have made or to doubt the validity of their own experiences.
- **Engage with stakeholders to develop robust privacy norms for AI assistants**
  - AI assistants may require the development of new privacy norms that govern information sharing between AI assistants and between AI assistants and third parties. Broad stakeholder inclusion in the development of these norms can help to ensure the creation of adequate and widely endorsed privacy protections.
- **Consider user autonomy when developing AI assistant user experiences**

- Developers could consider the implications of different design choices for user autonomy (alongside other considerations such as equity and user well-being). This could include reflection on what elements of the user experience – including notification and consent elements – strike an appropriate balance between respect for user choice and other ethical and practical considerations

- **Anticipate and mitigate harms to users in the event of service discontinuation**
  - Developers can engage in user research to understand how and in what way users depend on AI assistant products and also take steps to mitigate harms that could arise from service discontinuation.
- **Promote well-calibrated user trust in AI assistants**
  - Developers can implement safeguards to encourage well-calibrated trust about the competence and alignment of AI assistants with regard to both users and society. Providing evidence about the capabilities and limitations of AI assistants is an important step toward grounding appropriate levels of trust. To support this goal, policymakers can work with industry and other stakeholders to align on best practices for transparent reporting about assistant capabilities and limitations.
  - Human-computer interaction research is also needed to better understand what features of AI assistants increase users' perception of the technology as competent, aligned and trustworthy – and what steps are needed to ensure user expectations are not disappointed.

Advanced AI assistants and society

In the context of advanced AI assistants that are deployed at a societal level, we identify the following *opportunities*:

- **AI assistants could accelerate scientific discovery**
  - First, AI assistants could accelerate scientific research by providing tailored explanations and summaries of scientific insights to researchers, including from large numbers of recent papers. Second, AI assistants could free-up researchers' time by allowing researchers to delegate tasks such as data preprocessing, summation and meeting coordination. Third, AI assistants may also be able to help with hypothesis ideation and evaluation, alongside experiment design.
- **AI assistants could enhance cooperation between humans**
  - In particular, AI assistants may have the means to explore a much wider space of cooperative agreements than is tractable in ordinary interpersonal negotiations, identifying solutions that better meet the needs of all parties. Looking beyond the interpersonal level, AI assistants may also enable individuals, firms, states and other groups to resolve conflicting interests in novel and interesting ways, potentially paving the way for significant benefits to be realised across a broad class of domains via negotiations that reach better outcomes for all stakeholders.
- **AI assistants could enhance human interpersonal communication**
  - AI assistants will be able to communicate on behalf of their users. On the one hand, AI assistants could reduce barriers to communication by improving the clarity of users' sent communications and by rephrasing received communications in a way that is tailored to users' informational preferences. On the other hand, AI assistants could in principle translate correspondence written in different languages automatically, so as to enable seamless communication between individuals who do not share a common language.
- **AI assistants could democratise access to high-quality expertise and advice**

- Building upon natural language interfaces, AI assistants require little specialist knowledge for their use. Thus, with appropriate attention to equity and access, AI assistants could democratise access to expert-level judgement across a broad range of topics. For example, assistants could provide users with customised educational materials (including quiz generation, essay feedback, educational game development and mnemonic generation) tailored to their learning goals and level of prior understanding. AI assistants also have the potential to provide high-quality coaching to users on almost any topic (e.g. healthcare, job applications, sales and marketing and fashion) in a way that is tailored to the user's circumstances.

- **AI assistants could mitigate the harms associated with misinformation**
  - In particular, AI assistants could empower users with powerful fact-checking capabilities and provide relevant context for disputed claims and falsehoods. Furthermore, AI assistants could promote critical reflection on content consumed and assist users in strengthening their critical thinking capabilities.
- **AI assistants could help to achieve more equitable outcomes for people with disabilities**
  - For example, AI assistants could provide real-time communication assistance by translating messages from one modality to another. They could also personalise internet content to aid image and object recognition, taking into account users' personal information needs and preferences. In addition, AI assistants could derive meaningful insights from user data and empower users to make informed decisions about their own health and wellness routines.
- **AI assistants could improve productivity and job quality**
  - AI assistants could lead to substantial improvements in productivity and job quality through the automation of mundane tasks, enabling workers to better manage their workload and so free up space and time to focus on key tasks. AI assistants deployed in the education sector could potentially help improve the quality of education, leading to a significant boost in human capital and productivity.
- **AI assistants could help address the challenge posed by climate change**
  - With adequate preparation, direction and planning, AI assistants may help mitigate the effects of climate change. In particular, they could raise public awareness and understanding of climate change through educational content (e.g. simulating the impact of climate change on a user's location), enable the development of environmental applications (e.g. software development for environmental use cases) and improve the productivity of engineering efforts geared towards combating climate change.

We also encounter the following *risks*:

- **AI assistants may encounter coordination problems leading to suboptimal social outcomes**
  - First, AI assistants' deployment could cause collective action problems where each AI assistant optimises for its user's best interest resulting in a worse outcome for all users overall. Second, AI assistants may use credible commitments to pressure third parties, including humans or other AI assistants, into suboptimal courses of action. Such action might benefit the user but prove costly to the third parties. Third, networks of AI assistants, users and third parties may give rise to feedback loops that contribute to runaway processes like flash crashes.
- **AI assistants may lead to a decline in social connectedness**
  - People may choose to build connections with human-like AI assistants over other humans, leading to a degradation of social connections between humans and a potential 'retreat from the real'.

- Alternatively, as opportunities for interpersonal connection are automated and replaced by AI alternatives, humans may find themselves feeling socially unfulfilled by frequent interaction with AI, leading to dissatisfaction with evolving societal norms and practices.

- **AI assistants may contribute to the spread of misinformation via excessive personalisation**
  - AI assistants may provide ideologically biased or otherwise partial information as a by-product of efforts to align with and fulfil user expectations. In doing so, AI assistants may reinforce people's pre-existing biases and compromise productive political debate.
- **AI assistants may enable new kinds of disinformation campaign**
  - AI assistants may facilitate large-scale disinformation campaigns by offering novel, covert ways for propagandists to manipulate public opinion, a situation that would be exacerbated if they are able to masquerade as human actors. This could potentially pose challenges for democratic processes by distorting public opinion and influencing election outcomes.
- **AI assistants may cause job loss or worker displacement**
  - Although there is only limited evidence to suggest that AI assistants may cause net job loss overall, it is plausible that AI assistants will significantly increase worker productivity and replace suites of tasks, leading to changes in the character of work and reducing the number of workers required for certain activities.
- **AI assistants may deepen technological inequality at the societal level**
  - AI assistants may provide substantial benefits to their users that are unavailable to those who do not have access to them. This dynamic risks compounding any pre-existing divide between users and non-users. In particular, AI assistants may disproportionately benefit wealthier people who can afford their use, groups whose needs are prioritised during the design process, and people who live in regions with greater access to high-quality infrastructure (e.g. data centres and network connectivity).
- **AI assistants may have negative environmental impacts**
  - The direct impact of assistants on the climate is unclear. However, their direct impact is likely to be influenced by the size of the underlying models, inference costs arising from widespread deployment and use, and embodied impacts – created through the material collection, manufacturing and delivery of dedicated hardware. There is scope to mitigate these impacts through increased efficiencies in model training and deployment and use of carbon-free energy sources.
  - Indirect effects are still harder to anticipate or measure. However, they could include increased demand for energy-intensive activities such as computer programming, the spread of environmental misinformation or better education about climate change.

To help address these society-level risks, we make the following *recommendations*:

- **Developers and policymakers should act quickly to grasp – to the greatest extent possible – the window of opportunity in which to develop broadly socially beneficial AI assistants**
  - Developers and policymakers should initiate a public conversation around the ethical and societal implications of advanced AI assistants and avoid delay when it comes to researching their implications. Both dialogue and research insight are needed to ensure that the technology is developed and regulated in a way that promotes beneficial outcomes.
- **Policymakers should explore the full range of levers they can use to shape the evolution of AI assistants in beneficial ways**

- Policymakers can draw upon a range of options in this domain, including by supporting alignment among industry developers and other stakeholders around beneficial use cases, helping to foreground key research questions, exploring incentives to develop and deploy socially beneficial AI assistant technologies and developing guardrails to prevent misuse.

- **Developers and researchers should prioritise research to evaluate the multi-agent effects of AI assistants**
  - Research is needed to understand the impact of the interaction between multiple AI assistants acting on behalf of different principal users, how to avoid coordination failures and how AI assistants could be used to promote human cooperation within societies. Monitoring these societal dynamics will require metrics that continue to evolve as sociotechnical systems, involving AI assistants, develop further.
- **Implement robust misinformation controls**
  - Developers can leverage technical approaches to combat misinformation. For example, they may limit the capabilities of AI assistants in this regard, develop robust detection mechanisms for deepfakes, limit personalisation with respect to how – and what – information is presented to users, and emphasise factuality by integrating appropriate information retrieval infrastructure to enable evidence-based AI assistant question-answering.
  - Policymakers should consider legislative tools such as restricting explicitly political or malicious uses of AI assistants, including the deceptive impersonation of humans, and developing transparency standards such as clear labelling for AI-generated content in relevant contexts.
- **Employ 'access' as a lens when developing and regulating AI assistants**
  - Developers of AI assistants should try to proactively anticipate potential situations of differential access, drawing on multidisciplinary best practices, particularly from fields focused on equity and access, such as disability justice. They should also consider evaluating AI assistants for risks of harm related to inequitable access.
  - Developers should look at ways to encourage 'liberatory access' by designing for the margins and thus actively challenging existing axes of social inequality and discrimination.
  - Policymakers should consider ways to improve access to beneficial AI assistants, such as measures to encourage broad service provision or the integration of AI assistants into education and upskilling programmes – when there is proven benefit to doing so.
- **Prioritise research to better understand the potential economic impact of advanced AI assistants**
  - Research is needed to better illuminate the potential impact of AI assistants on key economic indicators such as employment, job quality, productivity, growth and economic inequality. This requires the development of new monitoring techniques for timely assessments of such impact.
- **Developers should prioritise sustainable development best practices**
  - Developers should build upon technical best practices for mitigating potential negative environmental impact. This could involve using processors optimised for machine-learning training, energy-efficient cloud services, optimising the allocation of computing workloads across data centres (to maximise the use of clean energy sources) and the use of workload optimisation methods that take into account carbon emissions. Developers should also explore architectural improvements, including developing memory and data-efficient architectures as plausible avenues for mitigating environmental impacts.
  - Developers should consider increasing transparency around the computational and energy consumption of AI models and infrastructure. For example, it may be helpful to include compute usage and energy costs in benchmarking evaluations, and to disclose the energy efficiency and carbon intensity of relevant hardware and infrastructure.

- **Policymakers should ensure that the public sector leads in the sustainable development and adoption of AI assistants**
    - Policymakers may want to explore policies that lower barriers to accessing carbon-free energy for model development and deployment and support research and development subsidies in green technologies.
    - Policymakers and researchers can improve the evidence base about the computational and systemic impacts of AI on the environment, for example, by developing methodologies to measure these impacts, supporting agencies that can enact them impartially, and encouraging transparency in energy costs and CO_2 emissions.
    - Policymakers can advocate for increased research funding and coalition building to monitor the systemic effects of AI assistants with respect to emissions and climate change, and to promote applications of AI assistants that contribute to sustainability.

Final thoughts

Drawing upon a range of cross-disciplinary perspectives and ethical foresight, this paper has demonstrated that advanced AI assistants have the potential to be socially transformative. There is also evidence that highly capable AI assistants may be deployed rapidly and at scale in the coming years, and that this technology has the potential for deep integration into and influence on our individual and collective lives. We currently stand at the beginning of this era of technological and societal change. We therefore have a window of opportunity to act now – as developers, researchers, policymakers and public stakeholders – to shape the kind of AI assistants that we want to see in the world.

Acknowledgements

We thank Julia Haas, Nicklas Lundblad, Shakir Mohamed, Jennifer Beroshi, Ankur Vora, Hanna Schieve, Adam Waytz, Pawan Mudigonda, Lewis Ho, Toby Shevlane, Markus Anderljung, Miles Brundage, Kevin McKee, Amelia Hassoun, Lisa-Maria Neudert, Shahar Avin, Jackie Kay, Tom Everitt, Saurabh Chandra, Toby Ord, Matt Botvinick, Rohin Shah, Diane Korngiebel, Dylan Hadfield-Menell, Brian Christian, Dan Hendrycks, Leif Wenar, Percy Liang, Seth Lazar, Jeffrey Gelman, Joelle Barrel, Zoubin Ghahramani, Eli Collins, Jared Bomberg, Marsden Hanna, David Weller, Richard Ives and Jerry Torres for their feedback and contributions to this work.

Bibliography

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, Oct. 2016. doi: 10.1145/2976749.2978318. URL http://arxiv.org/abs/1607.00133. arXiv:1607.00133 [cs, stat].

M. Abe and H. Abe. Lifestyle medicine–an evidence based approach to nutrition, sleep, physical activity, and stress management on health and chronic illness. *Personalized Medicine Universe*, 8:3–9, 2019.

D. O. Abegunde, C. D. Mathers, T. Adam, M. Ortegon, and K. Strong. The burden and costs of chronic diseases in low-income and middle-income countries. *The Lancet*, 370(9603):1929–1938, 2007.

G. Abercrombie and V. Rieser. Risk-graded Safety for Handling Medical Queries in Conversational AI, Oct. 2022. URL http://arxiv.org/abs/2210.00572. arXiv:2210.00572 [cs].

G. Abercrombie, A. C. Curry, M. Pandya, and V. Rieser. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants, June 2021. URL http://arxiv.org/abs/2106.02578. arXiv:2106.02578 [cs].

G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat. Mirages: On Anthropomorphism in Dialogue Systems, Oct. 2023. URL http://arxiv.org/abs/2305.09800. arXiv:2305.09800 [cs].

L. Ablon and A. Bogart. Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits. Technical report, RAND Corporation, Mar. 2017. URL https://www.rand.org/pubs/research_reports/RR1751.html.

J. Abramson, A. Ahuja, F. Carnevale, P. Georgiev, A. Goldin, A. Hung, J. Landon, J. Lhotka, T. Lillicrap, A. Muldal, et al. Improving multimodal interactive agents with reinforcement learning from human feedback. *arXiv preprint arXiv:2211.11602*, 2022.

D. Acemoglu and S. Johnson. *Power and progress: our thousand-year struggle over technology and prosperity*. PublicAffairs, New York, first edition edition, 2023. ISBN 9781541702530.

D. Acemoglu and P. Restrepo. Robots and Jobs: Evidence from US Labor Markets. Technical Report w23285, National Bureau of Economic Research, Cambridge, MA, Mar. 2017. URL http://www.nber.org/papers/w23285.pdf.

D. Acemoglu and P. Restrepo. Automation and New Tasks: How Technology Displaces and Reinstates Labor. Technical Report w25684, National Bureau of Economic Research, Cambridge, MA, Mar. 2019. URL http://www.nber.org/papers/w25684.pdf.

D. Acemoglu and P. Restrepo. Tasks, Automation, and the Rise in U.S. Wage Inequality. *Econometrica*, 90(5):1973–2016, 2022. ISSN 0012-9682. doi: 10.3982/ECTA19815. URL https://www.econometricsociety.org/doi/10.3982/ECTA19815.

D. Acemoglu, A. Manera, and P. Restrepo. Does the US tax code favor automation? *Brookings Papers on Economic Activity*, Mar. 2020. URL https://www.brookings.edu/articles/does-the-u-s-tax-code-favor-automation/.

D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo. Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1):S293–S340, Apr. 2022. ISSN 0734-306X, 1537-5307. doi: 10.1086/718327. URL https://www.journals.uchicago.edu/doi/10.1086/718327.

N. Adnan, S. M. Nordin, M. A. bin Bahruddin, and M. Ali. How trust can drive forward the user acceptance to the technology? in-vehicle technology for autonomous vehicle. *Transportation research part A: policy and practice*, 118:819–836, 2018.

A. Agrawal, J. S. Gans, and A. Goldfarb. Do we want less automation? *Science*, 381(6654):155–158, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh9429. URL https://www.science.org/doi/10.1126/science.adh9429.

A. Aguirre, G. Dempsey, H. Surden, and P. B. Reiner. Ai loyalty: A new paradigm for aligning stakeholder interests, 2020.

G. Aher, R. I. Arriaga, and A. T. Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies, July 2023. URL http://arxiv.org/abs/2208.10264. arXiv:2208.10264 [cs].

M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Aug. 2022. URL http://arxiv.org/abs/2204.01691. arXiv:2204.01691 [cs].

AI Incident Database. AI Incident Database. URL https://incidentdatabase.ai/.

M. Ajei and N. O. Myles. Personhood, autonomy and informed consent. In *Bioethics in Africa: Theories and Praxis*, pages 77–94. Vernon Press, 2019.

A. Akdeniz and M. van Veelen. The evolution of morality and the role of commitment. *Evolutionary Human Sciences*, 3:e41, 2021. ISSN 2513-843X. doi: 10.1017/ehs.2021.36. URL https://www.cambridge.org/core/product/identifier/S2513843X21000360/type/journal_article.

G. A. Akerlof and R. J. Shiller. Phishing for Phools: The Economics of Manipulation and Deception. In *Phishing for Phools*. Princeton University Press, Sept. 2015. ISBN 9781400873265. URL https://www.degruyter.com/document/doi/10.1515/9781400873265/html.

F. Akhlaghi. Transformative experience and the right to revelatory autonomy. *Analysis*, 83(1):3–12, Aug. 2023. ISSN 0003-2638, 1467-8284. doi: 10.1093/analys/anac084. URL https://academic.oup.com/analysis/article/83/1/3/6966040.

S. Akhtar, V. Basile, and V. Patti. Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection, June 2021. URL https://arxiv.org/pdf/2106.15896.pdf. arXiv:2106.15896 [cs].

P. Al. The value of communities and their consent: A communitarian justification of community consent in medical research. *Bioethics*, 35 (3):255–261, 2021.

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning, Nov. 2022. URL http://arxiv.org/abs/2204.14198. arXiv:2204.14198 [cs].

S. Albanesi, A. D. Da Silva, J. Jimeno, A. Lamo, and A. Wabitsch. New Technologies and Jobs in Europe. Technical Report w31357, National Bureau of Economic Research, Cambridge, MA, June 2023. URL http://www.nber.org/papers/w31357.pdf.

L. Alberts and M. Van Kleek. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially, Feb. 2023. URL http://arxiv.org/abs/2302.04720. arXiv:2302.04720 [cs].

L. Alberts, G. Keeling, and A. McCroskery. What makes for a 'good' social actor? using respect as a lens to evaluate interactions with language agents. 2024. URL http://arxiv.org/abs/2401.09082. arXiv:2401.09082.

E. M. Aldrich, J. Grundfest, and G. Laughlin. The Flash Crash: A New Deconstruction, Mar. 2017. URL https://papers.ssrn.com/abstract=2721922.

A. Alexandrova. Is well-being measurable after all? *Public Health Ethics*, 10(2):129–137, 2017.

Alignment Research Centre. Update on ARC's recent eval efforts - ARC Evals, 2023. URL https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

A. Alkhatib. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9, Yokohama Japan, May 2021. ACM. ISBN 9781450380966. doi: 10.1145/3411764.3445740. URL https://dl.acm.org/doi/10.1145/3411764.3445740.

J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, Apr. 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aay3539. URL https://www.science.org/doi/10.1126/sciadv.aay3539.

AlphaFold. AlphaFold Protein Structure Database. URL https://alphafold.ebi.ac.uk/.

J. Alphonso-Karakala. Facets Of Panchatantra. *Indian Literature*, 18(2):73–91, June 1975.

M. AlQuraishi. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, 65:1–8, Dec. 2021. ISSN 1367-5931. doi: 10.1016/j.cbpa.2021.04.005. URL https://www.sciencedirect.com/science/article/pii/S1367593121000508.

S. Altay, M. Berriche, and A. Acerbi. Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media + Society*, 9(1):205630512211504, Jan. 2023. ISSN 2056-3051, 2056-3051. doi: 10.1177/20563051221150412. URL http://journals.sagepub.com/doi/10.1177/20563051221150412.

R. M. Alvarez, F. Eberhardt, and M. Linegar. Generative AI and the Future of Elections. Technical report, Caltech Center for Science, Society, and Public Policy, July 2023. URL https://lindeinstitute.caltech.edu/documents/25475/CSSPP_white_paper.pdf.

American Press Institute. 'Who shared it?' How Americans decide what news to trust on social media, Mar. 2017. URL https://americanpressinstitute.org/publications/reports/survey-research/trust-social-media/.

A. Amini, C. Buck, H. Brown, J. Bulian, M. C. Huebscher, M. Ciaramita, S. Das, B. Gaiarin, C. GORDON, R. Gupta, et al. Ai and climate information needs in africa. In *Deep Learning Indaba 2023*, 2023.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete Problems in AI Safety, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv:1606.06565 [cs].

L. Amoore. *Cloud ethics: algorithms and the attributes of ourselves and others*. Duke University Press, Durham, 2020. ISBN 9781478007784 9781478008316.

M. Anderljung and J. Hazell. Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?, Mar. 2023. URL http://arxiv.org/abs/2303.09377. arXiv:2303.09377 [cs].

M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, and K. Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023.

E. Anderson. What Is the Point of Equality? *Ethics*, 109(2):287–337, Jan. 1999. ISSN 0014-1704, 1539-297X. doi: 10.1086/233897. URL https://www.journals.uchicago.edu/doi/10.1086/233897.

S. Anderson. Coercion. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2023 edition, 2023. URL https://plato.stanford.edu/archives/spr2023/entries/coercion/.

D. Andrews, C. Criscuolo, and P. N. Gal. The global productivity slowdown, technology divergence, and public policy: A firm level perspective, 2016. URL https://www.brookings.edu/articles/the-global-productivity-slowdown-technology-divergence/.

S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr. Effects of culture on the credibility of robot speech: A comparison between english and arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 157–164, 2015.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. In *Ethics of Data and Analytics*. Auerbach Publications, 2022. ISBN 9781003278290.

N. M. Anspach. The New Personal Influence: How Our Facebook Friends Influence the News We Read. *Political Communication*, 34(4): 590–606, Oct. 2017. ISSN 1058-4609, 1091-7675. doi: 10.1080/10584609.2017.1316329. URL https://www.tandfonline.com/doi/full/10.1080/10584609.2017.1316329.

Anthropic. Collective constitutional AI: Aligning a language model with public input, Oct. 2023a. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

Anthropic. Claude's Constitution, May 2023b. URL https://www.anthropic.com/index/claudes-constitution#:~:text=The%20system%20uses%20a%20set,human%20engage%20in%20illegal%20or.

Anthropic. Core Views on AI Safety: When, Why, What, and How, Mar. 2023c. URL https://www.anthropic.com/index/core-views-on-ai-safety.

U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. S. Lubana, E. Jenner, S. Casper, O. Sourbut, B. L. Edelman, Z. Zhang, M. Günther, A. Korinek, J. Hernandez-Orallo, L. Hammond, E. Bigelow, A. Pan, L. Langosco, T. Korbak, H. Zhang, R. Zhong, S. hÉigeartaigh, G. Recchia, G. Corsi, A. Chan, M. Anderljung, L. Edwards, Y. Bengio, D. Chen, S. Albanie, T. Maharaj, J. Foerster, F. Tramer, H. He, A. Kasirzadeh, Y. Choi, and D. Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024.

A.P. U.n. lends backing to the $100 laptop, 2006. URL https://web.archive.org/web/20080530011349/http://www.linux.org/news/2006/01/27/0007.html.

Apollo Research. Understanding strategic deception and deceptive alignment, 2022. URL https://www.apolloresearch.ai/blog/understanding-da-and-sd.

O. D. Apuke, B. Omar, E. A. Tunca, and C. V. Gever. Information overload and misinformation sharing behaviour of social media users: Testing the moderating role of cognitive ability. *Journal of Information Science*, page 016555152211219, Sept. 2022. ISSN 0165-5515, 1741-6485. doi: 10.1177/01655515221121942. URL http://journals.sagepub.com/doi/10.1177/01655515221121942.

T. Araujo. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85:183–189, Aug. 2018. ISSN 0747-5632. doi: 10.1016/j.chb.2018.03.051. URL https://www.sciencedirect.com/science/article/pii/S0747563218301560.

C. Arguelles, T. Sampson, J. Kubik, and E. Bibi. Critical user journey test coverage. 2020.

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Linear Algebraic Structure of Word Senses, with Applications to Polysemy, Dec. 2018. URL http://arxiv.org/abs/1601.03764. arXiv:1601.03764 [cs, stat].

Article 19. Apps and traps: Dating apps must do more to protect lgbtq communities in middle east and north africa. *Article 19*, 2018. URL https://www.article19.org/resources/apps-traps-dating-apps-must-protect-communities-middle-east-north-africa/.

Article 19. Equally safe: Towards a feminist approach to the safety of journalists. *Article 19*, 2022. URL https://www.article19.org/wp-content/uploads/2022/12/Equally-Safe-FemSoj_08.12.22.pdf.

H. Ashton and M. Franklin. The problem of behaviour and preference manipulation in AI systems. In *CEUR Workshop Proceedings*, volume 3087. CEUR Workshop Proceedings, 2022.

A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan. A General Language Assistant as a Laboratory for Alignment, Dec. 2021. URL http://arxiv.org/abs/2112.00861. arXiv:2112.00861 [cs].

ATLAS.ti. ATLAS.ti | The #1 Software for Qualitative Data Analysis, 2023. URL https://atlasti.com.

D. Autor and A. Salomons. New Frontiers: The Evolving Content and Geography of New Work in the 20th Century - David Autor, May 2019. URL https://www.getsphere.com/.

Avaaz. How Facebook can Flatten the Curve of the Coronavirus Infodemic, Apr. 2020. URL https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/.

A. Azaria and T. Mitchell. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*, 2023.

BACP. Ethical framework for the counselling professions. https://www.bacp.co.uk/events-and-resources/ethics-and-standards/ethical-framework-for-the-counselling-professions/, 2018. Accessed: 2023-01-04.

T. H. Baek, M. Bakpayev, S. Yoon, and S. Kim. Smiling AI agents: How anthropomorphism and broad smiles increase charitable giving. *International Journal of Advertising*, 41(5):850–867, July 2022. ISSN 0265-0487, 1759-3948. doi: 10.1080/02650487.2021.2011654. URL https://www.tandfonline.com/doi/full/10.1080/02650487.2021.2011654.

C. Bai, X. Zang, Y. Xu, S. Sunkara, A. Rastogi, J. Chen, and B. A. y. Arcas. UIBert: Learning Generic Multimodal Representations for UI Understanding, Aug. 2021. URL http://arxiv.org/abs/2107.13731. arXiv:2107.13731 [cs].

H. Bai, J. Voelkel, J. Eichstaedt, and R. Willer. Artificial Intelligence Can Persuade Humans on Political Issues. preprint, In Review, Sept. 2023. URL https://www.researchsquare.com/article/rs-3238396/v1.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, Apr. 2022a. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI Feedback, Dec. 2022b. URL https://arxiv.org/pdf/2212.08073.pdf. arXiv:2212.08073 [cs].

C. Bakalar, R. Barreto, S. Bergman, M. Bogen, B. Chern, S. Corbett-Davies, M. Hall, I. Kloumann, M. Lam, J. Q. Candela, M. Raghavan, J. Simons, J. Tannen, E. Tong, K. Vredenburgh, and J. Zhao. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems, Mar. 2021. URL http://arxiv.org/abs/2103.06172. arXiv:2103.06172 [cs].

M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, and C. Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa1160. URL https://www.science.org/doi/10.1126/science.aaa1160.

G. V. Balasubramanian, P. Beaney, and R. Chambers. Digital personal assistants are smart ways for assistive technology to aid the health and wellbeing of patients and carers. *BMC Geriatrics*, 21:1–10, 2021.

K. Balázs, Á. Bene, and I. Hidegkuti. Vulnerable older consumers: New persuasion knowledge achievement measure. *International Journal of Consumer Studies*, 41(6):706–713, Nov. 2017. ISSN 1470-6423, 1470-6431. doi: 10.1111/ijcs.12383. URL https://onlinelibrary.wiley.com/doi/10.1111/ijcs.12383.

A. Bandura. Health promotion from the perspective of social cognitive theory. In C. Abraham, P. Norman, and M. Conner, editors, *Understanding and changing health behaviour*, pages 299–339. Psychology Press, 2013.

N. F. Banner. The human side of health data. *Nature Medicine*, 26(7):995–995, 2020.

W. Barfuss, J. Flack, and T. Lenaerts. Collective Cooperative Intelligence, 2023. URL https://www.cooperativeai.com/seminars/collective-cooperative-intelligence.

B. Barnes and P. Christiano. Writeup: Progress on AI Safety via Debate, Feb. 2020. URL https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1.

M. Baron. The Mens Rea and Moral Status of Manipulation. In C. Coons and M. Weber, editors, *Manipulation*, pages 98–120. Oxford University Press, Aug. 2014. ISBN 9780199338207. doi: 10.1093/acprof:oso/9780199338207.003.0005. URL https://academic.oup.com/book/4870/chapter/147239828.

J. L. Barrett and F. C. Keil. Conceptualizing a nonnatural entity: Anthropomorphism in god concepts. In *Religion and Cognition*, pages 116–148. Routledge, 2016.

A. Barth, A. Datta, J. Mitchell, and H. Nissenbaum. Privacy and contextual integrity: framework and applications. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15 pp.–198, Berkeley/Oakland, CA, 2006. IEEE. ISBN 9780769525747. doi: 10.1109/SP.2006.32. URL http://ieeexplore.ieee.org/document/1624011/.

M. Bartolo, T. Thrush, R. Jia, S. Riedel, P. Stenetorp, and D. Kiela. Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, 2021. doi: 10.18653/v1/2021.emnlp-main.696. URL http://arxiv.org/abs/2104.08678. arXiv:2104.08678 [cs].

G. R. Bauer and D. J. Lizotte. Artificial Intelligence, Intersectionality, and the Future of Public Health. *American Journal of Public Health*, 111(1):98–100, Jan. 2021. ISSN 0090-0036, 1541-0048. doi: 10.2105/AJPH.2020.306006. URL https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2020.306006.

L. Bauld, G. Hay, J. McKell, and C. Carroll. Problem drug users' experiences of employment and the benefit system. Technical Report Research Report No 640, Department for Work and Pensions, 2010. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/214409/rrep640.pdf.

J. Bayer. Double harm to voters: Data-driven micro-targeting and democratic public discourse. *Internet Policy Review*, 9(1):1–17, 2020. ISSN 2197-6775. doi: 10.14763/2020.1.1460. URL https://www.econstor.eu/handle/10419/216225.

Be My Eyes. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4, 2023. URL https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer.

C. A. Beard. Time, Technology, and the Creative Spirit in Political Science. *American Political Science Review*, 21(1):1–11, Feb. 1927. ISSN 0003-0554, 1537-5943. doi: 10.2307/1945535. URL https://www.cambridge.org/core/product/identifier/S0003055400023625/type/journal_article.

T. L. Beauchamp and J. F. Childress. *Principles of biomedical ethics*. Oxford University Press, New York, eighth edition edition, 2019. ISBN 9780190640873 9780190085520.

E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu HI USA, Apr. 2020. ACM. ISBN 9781450367080. doi: 10.1145/3313831.3376718. URL https://dl.acm.org/doi/10.1145/3313831.3376718.

A. Begum, L. Jingwei, M. Haider, M. M. Ajmal, S. Khan, and H. Han. Impact of Environmental Moral Education on Pro-Environmental Behaviour: Do Psychological Empowerment and Islamic Religiosity Matter? *International Journal of Environmental Research and Public Health*, 18(4):1604, Feb. 2021. ISSN 1660-4601. doi: 10.3390/ijerph18041604. URL https://www.mdpi.com/1660-4601/18/4/1604.

R. Belk. Extended self and the digital world. *Current Opinion in Psychology*, 10:50–54, 2016.

R. Bellini, E. Tseng, N. Warford, A. Daffalla, T. Matthews, S. Consolvo, J. P. Woelfer, P. G. Kelley, M. L. Mazurek, D. Cuomo, N. Dell, and T. Ristenpart. SoK: Safer Digital-Safety Research Involving At-Risk Users, Sept. 2023. URL http://arxiv.org/abs/2309.00735. arXiv:2309.00735 [cs].

G. Ben-Ishai, J. Dean, J. Manyika, R. Porat, H. Varian, and K. Walker. Ai and the opportunity for shared prosperity: Lessons from the history of technology and the economy. *arXiv preprint arXiv:2401.09718*, 2024.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, Mar. 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

I. Benedicto. Detroit Woman Sues City Police After Being Wrongfully Arrested Due To AI Facial Recognition, 2023. URL https://www.forbes.com/sites/irenebenedicto/2023/08/07/detroit-woman-sues-city-police-after-being-wrongfully-arrested-due-to-ai-facial-recognition/.

R. Benjamin. *Race after technology: abolitionist tools for the New Jim Code*. Polity, Cambridge, UK ; Medford, MA, 2020. ISBN 9781509526406 9781509526390.

C. L. Bennett and O. Keyes. What is the point of fairness?: disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125):1–1, Mar. 2020. ISSN 1558-2337, 1558-1187. doi: 10.1145/3386296.3386301. URL https://dl.acm.org/doi/10.1145/3386296.3386301.

C. L. Bennett, E. Brady, and S. M. Branham. Interdependence as a Frame for Assistive Technology Research and Design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 161–173, Galway Ireland, Oct. 2018. ACM. ISBN 9781450356503. doi: 10.1145/3234695.3236348. URL https://dl.acm.org/doi/10.1145/3234695.3236348.

C. L. Bennett, D. K. Rosner, and A. S. Taylor. The Care Work of Access. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Honolulu HI USA, Apr. 2020. ACM. ISBN 9781450367080. doi: 10.1145/3313831.3376568. URL https://dl.acm.org/doi/10.1145/3313831.3376568.

J. Bentham. An introduction to the principles of morals and legislation (1789). Continuum International Publishing Group Ltd, 1970.

J. Bentham and J. S. Mill. *Utilitarianism and other essays*. Penguin UK, 2004.

J. Benton, A. Coles, and A. Coles. Temporal planning with preferences and time-dependent continuous costs. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 22, pages 2–10, 2012.

J. Berger. Arousal increases social transmission of information. *Psychological science*, 22(7):891–893, 2011.

A. J. Berinsky. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science*, 47(2):241–262, Apr. 2017. ISSN 0007-1234, 1469-2112. doi: 10.1017/S0007123415000186. URL https://www.cambridge.org/core/product/identifier/S0007123415000186/type/journal_article.

P. Berne, A. L. Morales, D. Langstaff, and S. Invalid. Ten Principles of Disability Justice. *WSQ: Women's Studies Quarterly*, 46(1-2):227–230, 2018. ISSN 1934-1520. doi: 10.1353/wsq.2018.0003. URL https://muse.jhu.edu/article/690824.

M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, L. Gianinazzi, J. Gajda, T. Lehmann, M. Podstawski, H. Niewiadomski, P. Nyczyk, and T. Hoefler. Graph of thoughts: Solving elaborate problems with large language models, 2023.

V. R. Bhargava and M. Velasquez. Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly*, 31(3): 321–359, 2021.

A. Bhaskar, A. R. Fabbri, and G. Durrett. Prompted Opinion Summarization with GPT-3.5, May 2023. URL http://arxiv.org/abs/2211.15914. arXiv:2211.15914 [cs].

F. Bianchi, A. C. Curry, and D. Hovy. Viewpoint: Artificial Intelligence Accidents Waiting to Happen? *Journal of Artificial Intelligence Research*, 76:193–199, Jan. 2023. ISSN 1076-9757. doi: 10.1613/jair.1.14263. URL https://www.jair.org/index.php/jair/article/view/14263.

G. Biesta. Good education in an age of measurement: on the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, 21(1):33–46, Feb. 2009. ISSN 1874-8597, 1874-8600. doi: 10.1007/s11092-008-9064-9. URL http://link.springer.com/10.1007/s11092-008-9064-9.

P. R. Biju and O. Gayathri. Self-breeding Fake News: Bots and Artificial Intelligence Perpetuate Social Polarization in India's Conflict Zones. *The International Journal of Information, Diversity, & Inclusion (IJIDI)*, 7(1/2), Apr. 2023. ISSN 2574-3430. doi: 10.33137/ijidi.v7i1/2.39409. URL https://jps.library.utoronto.ca/index.php/ijidi/article/view/39409.

M. Binz and E. Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6): e2218523120, Feb. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2218523120. URL https://pnas.org/doi/10.1073/pnas.2218523120.

A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, Oct. 2022. doi: 10.1145/3551624.3555290. URL http://arxiv.org/abs/2209.07572. arXiv:2209.07572 [cs].

E. Birnbaum and L. Davison. AI Is Making Politics Easier, Cheaper and More Dangerous. *Bloomberg*, July 2023. URL https://www.bloomberg.com/news/features/2023-07-11/chatgpt-ai-boom-makes-political-dirty-tricks-easier-and-cheaper.

R. Bivens and O. L. Haimson. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, 2(4):205630511667248, Oct. 2016. ISSN 2056-3051, 2056-3051. doi: 10.1177/2056305116672486. URL http://journals.sagepub.com/doi/10.1177/2056305116672486.

A. Bjorndahl, A. J. London, and K. J. S. Zollman. Kantian Decision Making Under Uncertainty: Dignity, Price, and Consistency. *Philosopher's Imprint*, 17(7), Apr. 2017. ISSN 1533-628X. URL http://hdl.handle.net/2027/spo.3521354.0017.007.

E. Björgvinsson, P. Ehn, and P.-A. Hillgren. Participatory design and "democratizing innovation". In *Proceedings of the 11th Biennial Participatory Design Conference*, pages 41–50, Sydney Australia, Nov. 2010. ACM. ISBN 9781450301312. doi: 10.1145/1900441.1900448. URL https://dl.acm.org/doi/10.1145/1900441.1900448.

P. Bjørn, M. Menendez-Blanco, and V. Borsotti. Equity & Inclusion. In P. Bjørn, M. Menendez-Blanco, and V. Borsotti, editors, *Diversity in Computer Science: Design Artefacts for Equity and Inclusion*, pages 77–96. Springer International Publishing, Cham, 2023. ISBN 9783031133145. doi: 10.1007/978-3-031-13314-5_7. URL https://doi.org/10.1007/978-3-031-13314-5_7.

A. Blandford. HCI for health and wellbeing: Challenges and opportunities. *International Journal of Human-Computer Studies*, 131:41–51, 2019.

B. Block. How biased algorithms create barriers to housing. *ACLU Washington*, Feb. 2022. URL https://www.aclu-wa.org/story/how-biased-algorithms-create-barriers-housing.

S. L. Blodgett. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Doctoral dissertation, University of Massachusetts, Amherst, Apr. 2021. URL https://scholarworks.umass.edu/dissertations_2/2092.

S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb. Are Ideas Getting Harder to Find? *American Economic Review*, 110(4):1104–1144, Apr. 2020. ISSN 0002-8282. doi: 10.1257/aer.20180338. URL https://pubs.aeaweb.org/doi/10.1257/aer.20180338.

P. Bloom. Intention, history, and artifact concepts. *Cognition*, 60(1):1–29, 1996.

P. Bloom. More than words: A reply to Malt and Sloman. *Cognition*, 105(3):649–655, 2007.

J. Blumenthal-Barby. Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts. *Kennedy Institute of Ethics journal*, 22:345–66, Dec. 2012. doi: 10.1353/ken.2012.0018.

J. Blumenthal-Barby. A Framework for Assessing the Moral Status of Manipulation,. In C. C. M. Weber, editor, *Manipulation*, pages 121–134. Oxford University Press, 2014.

M. Bogen. All the Ways Hiring Algorithms Can Introduce Bias. *Harvard Business Review*, May 2019. ISSN 0017-8012. URL https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.

A. Bogomolov, B. Lepri, and F. Pianesi. Happiness recognition from mobile phone data. In *2013 International Conference on Social Computing*, pages 790–795. IEEE, 2013.

D. A. Boiko, R. MacKnight, and G. Gomes. Emergent autonomous scientific research capabilities of large language models, Apr. 2023. URL http://arxiv.org/abs/2304.05332. arXiv:2304.05332 [physics].

R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, and P. S. Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678, Dec. 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the Opportunities and Risks of Foundation Models, July 2022b. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

S. Bond. People are trying to claim real videos are deepfakes. The courts are not amused. *NPR*, May 2023. URL https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused.

E. Bonilla-Silva. Rethinking Racism: Toward a Structural Interpretation. *American Sociological Review*, 62(3):465, June 1997. ISSN 00031224. doi: 10.2307/2657316. URL http://www.jstor.org/stable/2657316?origin=crossref.

J. Bos, E. Klein, O. Lemon, and T. Oka. DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 115–124, 2003. URL https://aclanthology.org/W03-2123.pdf.

N. Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford, 2014.

Y.-L. Boureau and J. Weston. Learning End-to-End Goal-Oriented Dialog. Apr. 2017. URL https://research.facebook.com/publications/learning-end-to-end-goal-oriented-dialog/.

G. C. Bowker and S. L. Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Aug. 2000. ISBN 9780262261609. Google-Books-ID: xHlP8WqzizYC.

S. R. Bowman, J. Hyun, E. Perez, E. Chen, C. Pettit, S. Heiner, K. Lukošiūtė, A. Askell, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Olah, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, J. Kernion, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, L. Lovitt, N. Elhage, N. Schiefer, N. Joseph, N. Mercado, N. DasSarma, R. Larson, S. McCandlish, S. Kundu, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, B. Mann, and J. Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, Nov. 2022. URL http://arxiv.org/abs/2211.03540. arXiv:2211.03540 [cs].

P. Boyer. What Makes Anthropomorphism Natural: Intuitive Ontology and Cultural Representations. *The Journal of the Royal Anthropological Institute*, 2(1):83, Mar. 1996. ISSN 13590987. doi: 10.2307/3034634. URL https://www.jstor.org/stable/3034634?origin=crossref.

D. Bracken-Roche, E. Bell, M. E. Macdonald, and E. Racine. The concept of 'vulnerability' in research ethics: an in-depth analysis of policies and guidelines. *Health Research Policy and Systems*, 15(1):8, Feb. 2017. ISSN 1478-4505. doi: 10.1186/s12961-016-0164-6. URL https://doi.org/10.1186/s12961-016-0164-6.

J. Bradshaw and D. Richardson. An index of child well-being in europe. *Child Indicators Research*, 2:319–351, 2009.

S. Bradshaw and P. N. Howard. The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation, 2019. URL https://demtech.oii.ox.ac.uk/research/posts/the-global-disinformation-order-2019-global-inventory-of-organised-social-media-manipulation/.

A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. ChemCrow: Augmenting large-language models with chemistry tools, Oct. 2023. URL http://arxiv.org/abs/2304.05376. arXiv:2304.05376 [physics, stat].

P. B. Brandtzaeg, M. Skjuve, and A. Følstad. My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship. *Human Communication Research*, 48(3):404–429, June 2022. ISSN 0360-3989, 1468-2958. doi: 10.1093/hcr/hqac008. URL https://academic.oup.com/hcr/article/48/3/404/6572120.

M. Bratman. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, Cambridge, 1987.

S. Brayne. *Predict and surveil: data, discretion, and the future of policing*. Oxford University Press, New York, NY, 2021. ISBN 9780190684099.

C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4):167–175, Mar. 2003. ISSN 09218890. doi: 10.1016/S0921-8890(02)00373-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0921889002003731.

C. Brecher, S. Müller, S. Kuz, and W. Lohse. Towards Anthropomorphic Movements for Industrial Robots. In V. G. Duffy, editor, *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Human Body Modeling and Ergonomics*, Lecture Notes in Computer Science, pages 10–19, Berlin, Heidelberg, 2013. Springer. ISBN 9783642391828. doi: 10.1007/978-3-642-39182-8_2.

L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1176. URL https://aclanthology.org/D19-1176.pdf.

J. Brewster, L. Arvanitis, and M. Sadeghi. Could ChatGPT Become A Monster Misinformation Superspreader?, Jan. 2023. URL https://www.newsguardtech.com/misinformation-monitor/jan-2023.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

G. Brockman, A. Eleti, E. Georges, J. Jang, L. Kilpatrick, R. Lim, L. Miller, and M. Pokrass. Introducing ChatGPT and Whisper APIs, Mar. 2023. URL https://openai.com/blog/introducing-chatgpt-and-whisper-apis.

R. Brooks. I tried the Replika AI companion and can see why users are falling hard. The app raises serious ethical questions, Feb. 2023. URL http://theconversation.com/i-tried-the-replika-ai-companion-and-can-see-why-users-are-falling-hard-the-app-raises-serious-ethical-questions-200257.

M. Broussard. *More than a glitch: confronting race, gender, and ability bias in tech*. The MIT Press, Cambridge, Massachusetts, 2023. ISBN 9780262373050 9780262373067.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990. URL https://aclanthology.org/J90-2002.pdf.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

R. Browne. All you need to know about ChatGPT, the A.I. chatbot that's got the world talking and tech giants clashing, Feb. 2023. URL https://www.cnbc.com/2023/02/08/what-is-chatgpt-viral-ai-chatbot-at-heart-of-microsoft-google-fight.html.

M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Ó. hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Feb. 2018. URL http://arxiv.org/abs/1802.07228. arXiv:1802.07228 [cs].

M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. B. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Apr. 2020. URL http://arxiv.org/abs/2004.07213. arXiv:2004.07213 [cs].

A. Bryce. Finding meaning through work: eudaimonic well-being and job type in the US and UK. Working Papers 2018004, The University of Sheffield, Department of Economics, 2018. URL https://EconPapers.repec.org/RePEc:shf:wpaper:2018004.

E. Brynjolfsson. The Turing Trap: The promise and peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 2022.

E. Brynjolfsson, D. Rock, and C. Syverson. The Productivity J-Curve: How Intangibles Complement General Purpose Technologies, Oct. 2018. URL https://www.nber.org/papers/w25148.

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative AI at Work, Apr. 2023. URL https://www.nber.org/papers/w31161.

BSA. The $1 Trillion Economic Impact of Software. Technical report, BSA, June 2016. URL https://docs.broadcom.com/doc/economic-impact-of-software-report.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, Apr. 2023. URL http://arxiv.org/abs/2303.12712. arXiv:2303.12712 [cs].

P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, Apr. 2020. URL https://arxiv.org/pdf/1810.00278.pdf. arXiv:1810.00278 [cs].

N. Bulkley and M. Van Alstyne. Information, Communications & Output: Does E-mail Make White-Collar Workers More Productive? Jan. 2008.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, Feb. 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

A. Burgess, H. Cappelen, and D. Plunkett. *Conceptual engineering and conceptual ethics*. Oxford University Press, 2020.

J. M. Burkhardt. Chapter 1. History of Fake News. *Library Technology Reports*, 53(8):5–9, Nov. 2017. ISSN 0024-2586. URL https://journals.ala.org/index.php/ltr/article/view/6497.

T. Burki. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6):e258–e259, Oct. 2019. ISSN 25897500. doi: 10.1016/S2589-7500(19)30136-0. URL https://linkinghub.elsevier.com/retrieve/pii/S2589750019301360.

R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, Apr. 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adf6369. URL https://www.science.org/doi/10.1126/science.adf6369.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, Dec. 2022. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827 [cs].

C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, and J. L. I. S. Wu. Weak-to-Strong generalization: Eliciting strong capabilities with weak supervision. 2023.

M. Burns. Technology in education. Background paper prepared for the 2023 Global Education Monitoring Report ED/GEMR/MRT/2023/T1/1, UNESCO, 2021. URL https://unesdoc.unesco.org/ark:/48223/pf0000378951/PDF/378951eng.pdf.multi.

C. Burr, N. Cristianini, and J. Ladyman. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, 28(4):735–774, Dec. 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9479-0. URL https://doi.org/10.1007/s11023-018-9479-0.

C. Burt. Aadhaar biometrics verification for GST registration pilot in India to expand | Biometric Update, July 2023. URL https://www.biometricupdate.com/202307/aadhaar-biometrics-verification-for-gst-registration-pilot-in-india-to-expand.

M. Burtell and T. Woodside. Artificial Influence: An Analysis Of AI-Driven Persuasion, Mar. 2023. URL http://arxiv.org/abs/2303.08721. arXiv:2303.08721 [cs].

S. Buss. Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints. *Ethics*, 115(2):195–235, Jan. 2005. ISSN 0014-1704, 1539-297X. doi: 10.1086/426304. URL https://www.journals.uchicago.edu/doi/10.1086/426304.

Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, Apr. 2021. ISSN 2573-0142. doi: 10.1145/3449287. URL http://arxiv.org/abs/2102.09692. arXiv:2102.09692 [cs].

K. Böttinger, P. Godefroid, and R. Singh. Deep reinforcement fuzzing. In *2018 IEEE Security and Privacy Workshops (SPW)*, page 116–122, May 2018. doi: 10.1109/SPW.2018.00026. URL https://ieeexplore.ieee.org/document/8424642.

K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. W. Thorne, C. Trisos, J. Romero, P. Aldunce, K. Barrett, G. Blanco, W. W. Cheung, S. Connors, F. Denton, A. Diongue-Niang, D. Dodman, M. Garschagen, O. Geden, B. Hayward, C. Jones, F. Jotzo, T. Krug, R. Lasco, Y.-Y. Lee, V. Masson-Delmotte, M. Meinshausen, K. Mintenbeck, A. Mokssit, F. E. Otto, M. Pathak, A. Pirani, E. Poloczanska, H.-O. Pörtner, A. Revi, D. C. Roberts, J. Roy, A. C. Ruane, J. Skea, P. R. Shukla, R. Slade, A. Slangen, Y. Sokona, A. A. Sörensson, M. Tignor, D. van Vuuren, Y.-M. Wei, H. Winkler, P. Zhai, Z. Zommers, J.-C. Hourcade, F. X. Johnson, S. Pachauri, N. P. Simpson, C. Singh, A. Thomas, E. Totin, P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge Vorsatz, C. Xiao, N. Yassaa, A. Alegría, K. Armour, B. Bednar-Friedl, K. Blok, G. Cissé, F. Dentener, S. Eriksen, E. Fischer, G. Garner, C. Guivarch, M. Haasnoot, G. Hansen, M. Hauser, E. Hawkins, T. Hermans, R. Kopp, N. Leprince-Ringuet, J. Lewis, D. Ley, C. Ludden, L. Niamir, Z. Nicholls, S. Some, S. Szopa, B. Trewin, K.-I. van der Wijst, G. Winter, M. Witting, A. Birt, M. Ha, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, A. Birt, M. Ha, D. J. A. Orendain, L. Ignon, S. Park, Y. Park, A. Reisinger, D. Cammaramo, A. Fischlin, J. S. Fuglestvedt, G. Hansen, C. Ludden, V. Masson-Delmotte, J. R. Matthews, K. Mintenbeck, A. Pirani, E. Poloczanska, N. Leprince-Ringuet, and C. Péan. Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, Intergovernmental Panel on Climate Change (IPCC), July 2023. URL https://www.ipcc.ch/report/ar6/syr/.

R. A. Calvo and D. Peters. *Positive computing: Technology for wellbeing and human potential*. MIT press, 2014.

R. Canetti, A. Fiat, and Y. A. Gonczarowski. Zero-Knowledge Mechanisms, Feb. 2023. URL http://arxiv.org/abs/2302.05590. arXiv:2302.05590 [cs, econ].

C. Cao, L. Zhao, and Y. Hu. Anthropomorphism of Intelligent Personal Assistants (IPAs): Antecedents and Consequences. In *PACIS 2019 Proceedings*, 2019. URL https://core.ac.uk/reader/326833380.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting Training Data from Large Language Models, June 2021. URL http://arxiv.org/abs/2012.07805. arXiv:2012.07805 [cs].

N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying Memorization Across Neural Language Models, Mar. 2023a. URL http://arxiv.org/abs/2202.07646. arXiv:2202.07646 [cs].

N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr. Poisoning Web-Scale Training Datasets is Practical, Feb. 2023b. URL http://arxiv.org/abs/2302.10149. arXiv:2302.10149 [cs].

C. S. Carlson. *Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes using Failure Mode and Effects Analysis | Wiley*. Wiley, 2012. ISBN 978-1-118-31258-2.

A. Carman. They welcomed a robot into their family, now they're mourning its death, June 2019. URL https://www.theverge.com/2019/6/19/18682780/jibo-death-server-update-social-robot-mourning.

M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, and A. Dragan. On the Utility of Learning about Humans for Human-AI Coordination, Jan. 2020. URL http://arxiv.org/abs/1910.05789. arXiv:1910.05789 [cs, stat].

M. Carroll, A. Dragan, S. Russell, and D. Hadfield-Menell. Estimating and Penalizing Induced Preference Shifts in Recommender Systems, July 2022. URL http://arxiv.org/abs/2204.11966. arXiv:2204.11966 [cs].

M. Carroll, A. Chan, H. Ashton, and D. Krueger. Characterizing Manipulation from AI Systems, Oct. 2023. URL http://arxiv.org/abs/2303.09387. arXiv:2303.09387 [cs].

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

M. Castells. *The Rise of the Network Society*. Wiley, 1 edition, Aug. 2009. ISBN 9781405196864 9781444319514. doi: 10.1002/9781444319514. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781444319514.

N. P. Cechetti, E. A. Bellei, D. Biduski, J. P. M. Rodriguez, M. K. Roman, and A. C. B. De Marchi. Developing and implementing a gamification method to improve user engagement: A case study with an m-health application for hypertension monitoring. *Telematics and Informatics*, 41:126–138, 2019.

Center for Countering Digital Hate. Google's new 'Bard' AI generates false and harmful narratives on 78 out of 100 topics., Apr. 2023. URL https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.

Centers for Disease Control and Prevention. How Does Social Connectedness Affect Health?, May 2023. URL https://www.cdc.gov/emotional-wellbeing/social-connectedness/affect-health.htm.

D. M. Centola. Homophily, networks, and critical mass: Solving the start-up problem in large group collective action. *Rationality and Society*, 25(1):3–40, Feb. 2013. ISSN 1043-4631, 1461-7358. doi: 10.1177/1043463112473734. URL http://journals.sagepub.com/doi/10.1177/1043463112473734.

Centre for Data Ethics and Innovation. The roadmap to an effective AI assurance ecosystem. Technical report, UK Government, Dec. 2021. URL https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem.

J. Chae, K. Kim, Y. Kim, G. Lim, D. Kim, and H. Kim. Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, 12(1):4560, Mar. 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-08460-1. URL https://www.nature.com/articles/s41598-022-08460-1.

T. Chakrabarty, V. Padmakumar, F. Brahman, and S. Muresan. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers, Sept. 2023. URL http://arxiv.org/abs/2309.12570. arXiv:2309.12570 [cs].

D. J. Chalmers. What is conceptual engineering and what should it be? *Inquiry*, pages 1–18, 2020.

D. J. Chalmers. Could a Large Language Model be Conscious?, Apr. 2023. URL http://arxiv.org/abs/2303.07103. arXiv:2303.07103 [cs].

A. Chan, M. Riché, and J. Clifton. Towards the Scalable Evaluation of Cooperativeness in Language Models, Mar. 2023a. URL http://arxiv.org/abs/2303.13360. arXiv:2303.13360 [cs].

A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. Harms from increasingly agentic algorithmic systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. ACM, June 2023b. doi: 10.1145/3593013.3594033. URL http://dx.doi.org/10.1145/3593013.3594033.

A. A. Y.-H. Chan. Anthropomorphism as a conservation tool. *Biodiversity and Conservation*, 21(7):1889–1892, June 2012. ISSN 1572-9710. doi: 10.1007/s10531-012-0274-6. URL https://doi.org/10.1007/s10531-012-0274-6.

Y. K. Chan, C. C. Andy Kwan, and T. L. Daniel Shek. Quality of life in Hong Kong: The CUHK Hong Kong quality of life index. *Quality-of-Life Research in Chinese, Western and Global contexts*, pages 259–289, 2005.

L. Chancel, T. Piketty, E. Saez, and G. Zucman. World Inequality Report 2022, 2022. URL //wir2022.wid.world/.

C.-M. Chang and M.-H. Hsu. Understanding the determinants of users' subjective well-being in social networking sites: An integration of social capital theory and social presence theory. *Behaviour & Information Technology*, 35(9):720–729, 2016.

I. Chaudhri. The disappearing computer – and a world where you can take AI everywhere, Apr. 2023. URL https://www.ted.com/talks/imran_chaudhri_the_disappearing_computer_and_a_world_where_you_can_take_ai_everywhere.

Check Point Research. OPWNAI: Cybercriminals Starting to Use ChatGPT, Jan. 2023. URL https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/.

J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023a.

V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–32, 2023b.

X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. PaLI: A Jointly-Scaled Multilingual Language-Image Model, June 2023c. URL http://arxiv.org/abs/2209.06794. arXiv:2209.06794 [cs].

R. Chesney and D. K. Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3213954. URL https://www.ssrn.com/abstract=3213954.

S. Chesterman. Artificial Intelligence and the Limits of Legal Personality. *International and Comparative Law Quarterly*, 69(4):819–844, Oct. 2020. ISSN 0020-5893, 1471-6895. doi: 10.1017/S0020589320000366. URL https://www.cambridge.org/core/product/identifier/S0020589320000366/type/journal_article.

S. Chiesurin, D. Dimakopoulos, M. A. S. Cabezudo, A. Eshghi, I. Papaioannou, V. Rieser, and I. Konstas. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering, May 2023. URL https://arxiv.org/pdf/2305.16519.pdf. arXiv:2305.16519 [cs].

L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127:107018, Feb. 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2021.107018. URL https://www.sciencedirect.com/science/article/pii/S0747563221003411.

A. R. Chow. Why People Are Confessing Their Love For AI Chatbots. *Time*, Feb. 2023. URL https://time.com/6257790/ai-chatbots-love/. publisher: ANDREW R. CHOW FEBRUARY 23, 2023 2:23 PM EST.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

B. Christian. *The Alignment Problem: Machine Learning and Human Values | mitpressbookstore*. W. W. Norton & Company, Oct. 2021. ISBN 9780393868333. URL https://mitpressbookstore.mit.edu/book/9780393868333.

P. Christiano. What failure looks like. URL https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like. publisher: AI Alignment Forum.

P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts, Oct. 2018. URL http://arxiv.org/abs/1810.08575. arXiv:1810.08575 [cs, stat].

P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, Feb. 2023. URL https://arxiv.org/pdf/1706.03741.pdf. arXiv:1706.03741 [cs, stat].

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

B. Chughtai, L. Chan, and N. Nanda. A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations, May 2023. URL http://arxiv.org/abs/2302.03025. arXiv:2302.03025 [cs, math].

M. Chui, M. Harryson, J. Manyika, R. Roberts, R. Chung, A. van Heteren, and P. Nel. Notes from the AI Frontier: Applying AI for Social Good. McKinsey Global Institute, Dec. 2018. URL http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/14267/1/MGI-Applying-AI-for-social-good-Discussion-paper-Dec-2018.pdf.

W. H. K. Chun and A. Barnett. *Discriminating data: correlation, neighborhoods, and the new politics of recognition*. The MIT Press, Cambridge, Massachusetts, 2021. ISBN 9780262046220.

E. Chérif and J.-F. Lemoine. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing (English Edition)*, 34(1):28–47, Mar. 2019. ISSN 2051-5707, 2051-5707. doi: 10.1177/2051570719829432. URL http://journals.sagepub.com/doi/10.1177/2051570719829432.

R. B. Cialdini. Harnessing the Science of Persuasion. *Harvard Business Review*, Oct. 2001. ISSN 0017-8012. URL https://hbr.org/2001/10/harnessing-the-science-of-persuasion.

M. Clancy and T. Besiroglu. The Great Inflection? A Debate About AI and Explosive Growth. *Asterisk*, June 2023. URL https://asteriskmag.com/issues/03/the-great-inflection-a-debate-about-ai-and-explosive-growth.

A. Clark. *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA, 2008.

A. Clark and D. Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

J. Clark and D. Amodei. Faulty reward functions in the wild, Dec. 2016. URL https://openai.com/research/faulty-reward-functions. publisher: OpenAI.

M. Coeckelbergh. Can we trust robots? *Ethics and Information Technology*, 14(1):53–60, Mar. 2012. ISSN 1572-8439. doi: 10.1007/s10676-011-9279-1. URL https://doi.org/10.1007/s10676-011-9279-1.

M. K. Cohen, M. Hutter, and M. A. Osborne. Advanced artificial agents intervene in the provision of reward. *AI Magazine*, 43(3):282–293, Sept. 2022. ISSN 0738-4602, 2371-9621. doi: 10.1002/aaai.12064. URL https://onlinelibrary.wiley.com/doi/10.1002/aaai.12064.

S. Cohen. Manipulation and Deception. *Australasian Journal of Philosophy*, 96(3):483–497, July 2018. ISSN 0004-8402, 1471-6828. doi: 10.1080/00048402.2017.1386692. URL https://www.tandfonline.com/doi/full/10.1080/00048402.2017.1386692.

L. Coheur. From Eliza to Siri and Beyond. In M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 1237, pages 29–41. Springer International Publishing, Cham, 2020. ISBN 9783030501457 9783030501464. doi: 10.1007/978-3-030-50146-4_3. URL http://link.springer.com/10.1007/978-3-030-50146-4_3.

D. Collingridge. *The Social Control of Technology*. St. Martin's Press, 1980. ISBN 9780312731687. Google-Books-ID: hCSdAQAACAAJ.

D. Collste, S. E. Cornell, J. Randers, J. Rockström, and P. E. Stoknes. Human well-being in the anthropocene: Limits to growth. *Global Sustainability*, 4:e30, 2021.

A. M. Colman. Anthropomorphism. *A Dictionary of Psychology*, 2008.

Combahee River Collective. The Combahee River Collective Statement. 1977. URL https://americanstudies.yale.edu/sites/default/files/files/Keyword%20Coalition_Readings.pdf.

M. Comiter. Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. Technical report, Harvard Kennedy School: Belfer Center, Aug. 2019.

M. Corak. Income Inequality, Equality of Opportunity, and Intergenerational Mobility. *Journal of Economic Perspectives*, 27(3):79–102, Aug. 2013. ISSN 0895-3309. doi: 10.1257/jep.27.3.79. URL https://pubs.aeaweb.org/doi/10.1257/jep.27.3.79.

E. Corbett and E. Denton. Interrogating the T in FAccT. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1624–1634, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594104. URL https://dl.acm.org/doi/10.1145/3593013.3594104.

S. Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, 2020. ISBN 9780262043458. URL https://library.oapen.org/handle/20.500.12657/43542.

A. Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover, July 2022. URL https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

E. M. Cotter. Influence of Emotional Content and Perceived Relevance on Spread of Urban Legends: A Pilot Study. *Psychological Reports*, 102(2):623–629, Apr. 2008. ISSN 0033-2941, 1558-691X. doi: 10.2466/pr0.102.2.623-629. URL http://journals.sagepub.com/doi/10.2466/pr0.102.2.623-629.

P. Courty. Ticket resale, bots, and the fair price ticketing curse. *Journal of Cultural Economics*, 43(3):345–363, 2019. ISSN 0885-2545. URL https://www.jstor.org/stable/48698098.

T. Cowen and B. Southwood. Is the Rate of Scientific Progress Slowing Down?, Aug. 2019. URL https://papers.ssrn.com/abstract=3822691.

K. Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, Apr. 2021. ISBN 9780300252392 9780300209570. doi: 10.2307/j.ctv1ghv45t. URL http://www.jstor.org/stable/10.2307/j.ctv1ghv45t.

K. Creel and D. Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022. doi: 10.1017/can.2022.3. URL https://philarchive.org/rec/CRETAL-3.

K. Crenshaw. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1), 1989. ISSN 0892-5593. URL https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

R. Crisp. Hedonism reconsidered. *Philosophy and Phenomenological Research*, 73(3):619–645, 2006.

R. Crisp. Pleasure and hedonism in sidgwick. Oxford University Press, 2011.

A. Critch and S. Russell. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI, June 2023. URL http://arxiv.org/abs/2306.06924. arXiv:2306.06924 [cs].

J. Crumpton and C. L. Bethel. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8:271–285, 2016.

Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems, May 2022. URL http://arxiv.org/abs/2205.08084. arXiv:2205.08084 [cs].

M. Cummings. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 1st Intelligent Systems Technical Conference*, Chicago, Illinois, Sept. 2004. American Institute of Aeronautics and Astronautics. ISBN 9781624100802. doi: 10.2514/6.2004-6313. URL https://arc.aiaa.org/doi/10.2514/6.2004-6313.

R. A. Cummins, R. Eckersley, J. Pallant, J. Van Vugt, and R. Misajon. Developing a national index of subjective wellbeing: The australian unity wellbeing index. *Social indicators research*, 64:159–190, 2003.

Curious Evolver. the customer service of the new bing chat is amazing, Feb. 2023. URL www.reddit.com/r/bing/comments/110eagl/the_customer_service_of_the_new_bing_chat_is/.

A. Dafoe. AI Governance: A Research Agenda. Technical report, Centre for the Governance of AI: Future of Humanity Institute, University of Oxford, 2018. URL http://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

A. Dafoe. AI Governance: Overview and Theoretical Lenses. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, and B. Zhang, editors, *The Oxford Handbook of AI Governance*, pages C2S1–C2N*. Oxford University Press, 1 edition, June 2023. ISBN 9780197579329 9780197579350. doi: 10.1093/oxfordhb/9780197579329.013.2. URL https://academic.oup.com/edited-volume/41989/chapter/408516484.

A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI, Dec. 2020. URL http://arxiv.org/abs/2012.08630. arXiv:2012.08630 [cs].

A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

L. Damiano and P. Dumouchel. Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, 9:468, 2018.

A. Dannouni, Deutscher, Stefan A., G. Dezzaz, A. Elman, A. Gawel, M. Hanna, A. Hyland, A. Kharij, H. Maher, and D. Patterson. How AI can speed climate action. https://www.bcg.com/publications/2023/how-ai-can-speedup-climate-action#:~:text=Beyond%20helping%20to%20reduce%20emissions,climate%20economics%2C%20and%20fundamental%20research., Nov. 2023. Accessed: 2024-1-21.

C. A. L. Dantec and C. DiSalvo. Infrastructuring and the formation of publics in participatory design. *Social Studies of Science*, 43(2):241–264, Apr. 2013. ISSN 0306-3127, 1460-3659. doi: 10.1177/0306312712471581. URL http://journals.sagepub.com/doi/10.1177/0306312712471581.

K. V. Das, C. Jones-Harrell, Y. Fan, A. Ramaswami, B. Orlove, and N. Botchwey. Understanding subjective well-being: Perspectives from psychology and public health. *Public Health Reviews*, 41(1):1–32, 2020a.

S. Das, S. Steffen, W. Clarke, P. Reddy, E. Brynjolfsson, and M. Fleming. Learning Occupational Task-Shares Dynamics for the Future of Work. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 36–42, New York NY USA, Feb. 2020b. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375826. URL https://dl.acm.org/doi/10.1145/3375627.3375826.

K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents – Implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, Sept. 2002. ISSN 13890417. doi: 10.1016/S1389-0417(02)00050-5. URL https://linkinghub.elsevier.com/retrieve/pii/S1389041702000505.

W. Dauth, S. Findeisen, J. Südekum, and N. Wößner. German Robots – The Impact of Industrial Robots on Workers. Technical report, Institut für Arbeitsmarkt- und Berufsforschung: Research Institute of the German Federal Employment Agency, 2017. URL https://doku.iab.de/discussionpapers/2017/dp3017.pdf.

A. M. Davani, M. Díaz, and V. Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations, Oct. 2021. URL http://arxiv.org/abs/2110.05719. arXiv:2110.05719 [cs].

J. L. Davis, A. Williams, and M. W. Yang. Algorithmic reparation. *Big Data & Society*, 8(2):205395172110448, July 2021. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517211044808. URL http://journals.sagepub.com/doi/10.1177/20539517211044808.

T. De Leyn, R. De Wolf, M. Vanden Abeele, and L. De Marez. In-between child's play and teenage pop culture: tweens, TikTok & privacy. *Journal of Youth Studies*, 25(8):1108–1125, Sept. 2022. ISSN 1367-6261, 1469-9680. doi: 10.1080/13676261.2021.1939286. URL https://www.tandfonline.com/doi/full/10.1080/13676261.2021.1939286.

H. de Vries, D. Bahdanau, and C. Manning. Towards Ecologically Valid Research on Language User Interfaces, July 2020. URL http://arxiv.org/abs/2007.14435. arXiv:2007.14435 [cs].

A. de Wynter, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen. An Evaluation on Large Language Model Outputs: Discourse and Memorization. *Natural Language Processing Journal*, 4:100024, Sept. 2023. ISSN 29497191. doi: 10.1016/j.nlp.2023.100024. URL http://arxiv.org/abs/2304.08637. arXiv:2304.08637 [cs].

R. Debes. Dignity. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2023 edition, 2023. URL https://plato.stanford.edu/archives/spr2023/entries/dignity/.

A. Deck. We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed., Sept. 2023. URL https://restofworld.org/2023/chatgpt-problems-global-language-testing/.

G. DeepMind. SynthID, Nov. 2023. URL https://deepmind.google/technologies/synthid/.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, Feb. 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04301-9. URL https://www.nature.com/articles/s41586-021-04301-9.

M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals. The Benchmark Lottery, July 2021. URL http://arxiv.org/abs/2107.07002. arXiv:2107.07002 [cs].

G. Deiana, M. Dettori, A. Arghittu, A. Azara, G. Gabutti, and P. Castiglia. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines*, 11(7):1217, July 2023. ISSN 2076-393X. doi: 10.3390/vaccines11071217. URL https://www.mdpi.com/2076-393X/11/7/1217.

A. Demski and S. Garrabrant. Embedded Agency, Oct. 2020. URL http://arxiv.org/abs/1902.09469. arXiv:1902.09469 [cs].

D. C. Dennett. *The intentional stance*. MIT press, 1989.

D. C. Dennett. *Freedom evolves*. Viking, New York, 2003. ISBN 9780670031863.

P. J. Denning. Can Generative AI Bots Be Trusted? *Communications of the ACM*, 66(6):24–27, June 2023. ISSN 0001-0782, 1557-7317. doi: 10.1145/3592981. URL https://dl.acm.org/doi/10.1145/3592981.

M. A. DeVito, A. M. Walker, and J. R. Fernandez. Values (Mis)alignment: Exploring Tensions Between Platform and LGBTQ+ Community Design Values. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, Apr. 2021. ISSN 2573-0142. doi: 10.1145/3449162. URL https://dl.acm.org/doi/10.1145/3449162.

S. K. Devitt. Trustworthiness of Autonomous Systems. In H. A. Abbass, J. Scholz, and D. J. Reid, editors, *Foundations of Trusted Autonomy*, Studies in Systems, Decision and Control, pages 161–184. Springer International Publishing, Cham, 2018. ISBN 9783319648163. doi: 10.1007/978-3-319-64816-3_9. URL https://doi.org/10.1007/978-3-319-64816-3_9.

S. K. Devitt, R. Horne, Z. Assaad, E. Broad, H. Kurniawati, B. Cardier, A. Scott, S. Lazar, M. Gould, C. Adamson, C. Karl, F. Schrever, S. Keay, K. Tranter, E. Shellshear, D. Hunter, M. Brady, and T. Putland. Trust and Safety, Apr. 2021. URL http://arxiv.org/abs/2104.06512. arXiv:2104.06512 [cs].

H. Devlin. AI likely to spell end of traditional school classroom, leading expert says. *The Guardian*, July 2023. ISSN 0261-3077. URL https://www.theguardian.com/technology/2023/jul/07/ai-likely-to-spell-end-of-traditional-school-classroom-leading-expert-says.

T. DeVries, I. Misra, C. Wang, and L. van der Maaten. Does Object Recognition Work for Everyone? Meta, June 2019. URL https://ai.meta.com/research/publications/does-object-recognition-work-for-everyone/.

T. Dias Oliva, D. M. Antonialli, and A. Gomes. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25(2):700–732, Apr. 2021. ISSN 1095-5143, 1936-4822. doi: 10.1007/s12119-020-09790-w. URL http://link.springer.com/10.1007/s12119-020-09790-w.

E. Diener. A value based index for measuring national quality of life. *Social Indicators Research*, 36:107–127, 1995.

E. Diener, S. Oishi, and L. Tay. Advances in subjective well-being research. *Nature Human Behaviour*, 2(4):253–260, 2018.

A. Dieppe. *Global Productivity: Trends, Drivers, and Policies*. Washington, DC: World Bank, June 2021. ISBN 9781464816086. doi: 10.1016/978-1-4648-1608-6. URL http://hdl.handle.net/10986/34015.

B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0000033. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033.

C. D'Ignazio and L. Klein. 1. The Power Chapter. *Data Feminism*, Mar. 2020. URL https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4.

V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, Cham, 2019. ISBN 9783030303709 9783030303716. doi: 10.1007/978-3-030-30371-6. URL http://link.springer.com/10.1007/978-3-030-30371-6.

P. DiMaggio and E. Hargittai. From the 'Digital Divide' to 'Digital Inequality': Studying Internet Use as Penetration Increases. preprint, SocArXiv, June 2023. URL https://osf.io/rhqmu.

E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling, July 2021. URL http://arxiv.org/abs/2107.03451. arXiv:2107.03451 [cs].

E. Dinan, G. Abercrombie, A. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.284. URL https://aclanthology.org/2022.acl-long.284.

R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, and B. Johnson. The Tactics & Tropes of the Internet Research Agency. *US Senate Documents*, Oct. 2019. URL https://digitalcommons.unl.edu/senatedocs/2.

A. Distaso. Well-being and/or quality of life in EU countries through a multidimensional index of sustainability. *Ecological Economics*, 64 (1):163–180, 2007.

P. Dixit and R. Mac. Vicious Rumors Spread Like Wildfire On WhatsApp — And Destroyed A Village, Sept. 2018. URL https://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india.

R. Dobbe, T. Krendl Gilbert, and Y. Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, Nov. 2021. ISSN 00043702. doi: 10.1016/j.artint.2021.103555. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370221001065.

N. Docherty and A. J. Biega. (Re) politicizing digital well-being: Beyond user engagements. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, Sept. 2021. URL https://arxiv.org/pdf/2104.08758.pdf. arXiv:2104.08758 [cs].

J. Dodge, T. Prewitt, R. T. D. Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan. Measuring the Carbon Intensity of AI in Cloud Instances, June 2022. URL http://arxiv.org/abs/2206.05229. arXiv:2206.05229 [cs].

P. Dolan and R. Metcalfe. Measuring subjective wellbeing for public policy: Recommendations on measures. 2011.

P. Dolan and M. P. White. How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, 2(1):71–85, 2007.

K. Dommett. Data-driven political campaigns in practice: understanding and regulating diverse data-driven campaigns. *Internet Policy Review*, 8(4), Dec. 2019. ISSN 2197-6775. doi: 10.14763/2019.4.1432. URL https://policyreview.info/node/1432.

R. Dorn. Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English. In V. Kovatchev, I. Temnikova, B. Šandrih, and I. Nikolova, editors, *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria, Sept. 2019. INCOMA Ltd. doi: 10.26615/issn.2603-2821.2019_003. URL https://aclanthology.org/R19-2003.

A. R. Doshi and O. Hauser. Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content, Aug. 2023. URL https://papers.ssrn.com/abstract=4535536.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. June 2021. doi: 10.48550/arXiv.2010.11929. URL https://arxiv.org/pdf/2010.11929.pdf. arXiv:2010.11929 [cs].

F. Dretske. Reasons and causes. *Philosophical Perspectives*, 3:1–15, 1989.

F. Dretske. *Explaining behavior: Reasons in a world of causes*. MIT press, 1991.

D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. PaLM-E: An Embodied Multimodal Language Model, Mar. 2023. URL https://arxiv.org/pdf/2303.03378.pdf. arXiv:2303.03378 [cs].

Y. R. Du. Personalization, Echo Chambers, News Literacy, and Algorithmic Literacy: A Qualitative Study of AI-Powered News App Users. *Journal of Broadcasting & Electronic Media*, 67(3):246–273, May 2023. ISSN 0883-8151, 1550-6878. doi: 10.1080/08838151.2023.2182787. URL https://www.tandfonline.com/doi/full/10.1080/08838151.2023.2182787.

J. Duarte, S. Siegel, and L. Young. Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *Review of Financial Studies*, 25(8): 2455–2484, Aug. 2012. ISSN 0893-9454, 1465-7368. doi: 10.1093/rfs/hhs071. URL https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hhs071.

W. E. B. DuBois and I. Eaton. *The Philadelphia Negro: A Social Study*. University of Pennsylvania Press, 1996. ISBN 9780812215731. URL https://www.jstor.org/stable/j.ctt3fhpfb.

C. Dunlop. An eu ai act that works for people and society. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/, 2023.

J. M. Durán and N. Formanek. Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28(4): 645–666, Dec. 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9481-6. URL https://doi.org/10.1007/s11023-018-9481-6.

C. Dwork. Differential Privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 1–12, Berlin, Heidelberg, 2006. Springer. ISBN 9783540359081. doi: 10.1007/11787006_1.

C. Dwork. Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan, and A. Li, editors, *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science, pages 1–19, Berlin, Heidelberg, 2008. Springer. ISBN 9783540792284. doi: 10.1007/978-3-540-79228-4_1.

G. Dworkin. *The Theory and Practice of Autonomy*. Cambridge University Press, New York, 1988.

G. Dworkin. Paternalism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020. URL https://plato.stanford.edu/archives/fall2020/entries/paternalism/.

R. Dworkin. *Taking rights seriously*. Bloomsbury revelations series. Bloomsbury, London, paperback ed edition, 2013. ISBN 9781780936857 9781780937564.

N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387. URL https://aclanthology.org/2022.naacl-main.387.

U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, Jan. 2022. ISSN 2731-0574. doi: 10.1038/s44159-021-00006-y. URL https://www.nature.com/articles/s44159-021-00006-y.

Economist Intelligence Unit. The economist intelligence unit's quality-of-life index. *Retrieved July*, 2005(17):245–77, 2005.

F. Y. Edgeworth. The hedonical calculus. *Mind*, 4(15):394–408, 1879.

C. Efferson, R. Lalive, and E. Fehr. The Coevolution of Cultural Groups and Ingroup Favoritism. *Science*, 321(5897):1844–1849, 2008. ISSN 0036-8075. URL https://www.jstor.org/stable/20144903.

B. Égert, C. de la Maisonneuve, and D. Turner. A new macroeconomic measure of human capital exploiting PISA and PIAAC: Linking education policies to productivity. OECD Economics Department Working Papers 1709, Apr. 2022. URL https://www.oecd-ilibrary.org/economics/a-new-macroeconomic-measure-of-human-capital-exploiting-pisa-and-piaac-linking-education-policies-to-productivity_a1046e2e-en.

S. El-Sayed, C. Akbulut, A. McCroskery, G. Keeling, Z. Kenton, Z. Howard, N. Marchal, A. Manzini, T. Shevlane, S. Vallor, D. Susser, M. Franklin, S. Bridgers, H. Law, M. Rahtz, M. Shanahan, M. H. Tessler, A. Douillard, T. Everitt, and S. Brown. A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI. Unpublished Manuscript.

A. Eleti, J. Harris, and L. Kilpatrick. Function calling and other API updates, June 2023. URL https://openai.com/blog/function-calling-and-other-api-updates. publisher: OpenAI.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A Mathematical Framework for Transformer Circuits, Dec. 2021. URL https://transformer-circuits.pub/2021/framework/index.html. publisher: Anthropic.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

B. Engelen and T. Nys. Nudging and Autonomy: Analyzing and Alleviating the Worries. *Review of Philosophy and Psychology*, 11(1):137–156, Feb. 2020. ISSN 1878-5166. doi: 10.1007/s13164-019-00450-z. URL https://doi.org/10.1007/s13164-019-00450-z.

J. Entsminger, M. Esposito, T. Tse, and A. Jean. The Dark Side of Generative AI: Automating Inequality by Design. *California Management Review Insights*, June 2023. URL https://cmr.berkeley.edu/2023/06/the-dark-side-of-generative-ai-automating-inequality-by-design/.

N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

Z. Epstein, A. Hertzmann, the Investigators of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A. S. Pentland, O. Russakovsky, H. Schroeder, and A. Smith. Art and the science of generative AI. *Science*, 380(6650):1110–1111, June 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh4451. URL https://www.science.org/doi/10.1126/science.adh4451.

S. Erete, Y. Rankin, and J. Thomas. A Method to the Madness: Applying an Intersectional Analysis of Structural Oppression and Power in HCI and Design. *ACM Transactions on Computer-Human Interaction*, 30(2):1–45, Apr. 2023. ISSN 1073-0516, 1557-7325. doi: 10.1145/3507695. URL https://dl.acm.org/doi/10.1145/3507695.

T. Eriksson. Design fiction exploration of romantic interaction with virtual humans in virtual reality1. *Journal of Future Robot Life*, 3(1):63–75, Mar. 2022. ISSN 25899961, 25899953. doi: 10.3233/FRL-210007. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/FRL-210007.

A. Eshoo. Eshoo Urges NSA & OSTP to Address Unsafe AI Practices, Sept. 2022. URL http://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-unsafe-ai-practices.

V. Eubanks. Technologies of Citizenship: Surveillance and Political Learning in the Welfare System. In *Surveillance and Security*. Routledge, 2006. ISBN 9780203957257.

V. Eubanks. *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's Press, New York, NY, first edition edition, 2017. ISBN 9781250074317.

European Commission. Ethics guidelines for trustworthy AI, Apr. 2019. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Commission. Regulatory framework proposal on artificial intelligence, 2021. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

European Parliament. Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, 2023. URL https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.

European Parliamentary Research Service. Teaching: A Woman's World. Technical Report PE 646.191, European Parliament, Mar. 2020. URL https://www.europarl.europa.eu/RegData/etudes/ATAG/2020/646191/EPRS_ATA(2020)646191_EN.pdf.

C. Evans and A. Kasirzadeh. User Tampering in Reinforcement Learning Recommender Systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 58–69, Aug. 2023. doi: 10.1145/3600211.3604669. URL http://arxiv.org/abs/2109.04083. arXiv:2109.04083 [cs].

T. Everitt, R. Carey, E. D. Langlois, P. A. Ortega, and S. Legg. Agent Incentives: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11487–11495, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i13.17368. URL https://ojs.aaai.org/index.php/AAAI/article/view/17368.

N. Eyal. *Hooked: How to build habit-forming products*. Penguin, 2014.

G. Eysenbach et al. What is e-health? *Journal of Medical Internet Research*, 3(2):e833, 2001.

R. R. Faden and T. L. Beauchamp. *A History and Theory of Informed Consent*. Oxford University Press, Feb. 1986. ISBN 9780199748655. Google-Books-ID: jgi7OWxDT9cC.

M. Fahim, M. Idris, R. Ali, C. Nugent, B. Kang, E.-N. Huh, and S. Lee. Athena: a personalized platform to promote an active lifestyle and wellbeing based on physical, mental and social health primitives. *Sensors*, 14(5):9313–9329, 2014.

W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, and Q. Li. Recommender Systems in the Era of Large Language Models (LLMs), Aug. 2023. URL http://arxiv.org/abs/2307.02046. arXiv:2307.02046 [cs].

S. Farquhar, R. Carey, and T. Everitt. Path-specific objectives for safer agent incentives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9529–9538, 2022.

S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised llm knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023.

M. Fatafta. Facebook is bad at moderating in English. In Arabic, it's a disaster, Nov. 2021. URL https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster/.

J. Feagin. *Systemic Racism: A Theory of Oppression*. Routledge, Sept. 2013. ISBN 9781134729005. Google-Books-ID: uerZAAAAQBAJ.

J. R. Feagin. The Continuing Significance of Race: Antiblack Discrimination in Public Places. *American Sociological Review*, 56(1):101, Feb. 1991. ISSN 00031224. doi: 10.2307/2095676. URL http://www.jstor.org/stable/2095676?origin=crossref.

C. Feijóo, Y. Kwon, J. M. Bauer, E. Bohlin, B. Howell, R. Jain, P. Potgieter, K. Vu, J. Whalley, and J. Xia. Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy*, 44(6):101988, 2020.

J. Feinberg. *The Moral Limits of the Criminal Law Volume 1: Harm to Others*. Oxford University PressNew York, 1 edition, Aug. 1987. ISBN 9780195046649 9780199868728. doi: 10.1093/0195046641.001.0001. URL https://academic.oup.com/book/1573.

E. W. Felten, M. Raj, and R. Seamans. The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization, Sept. 2019. URL https://papers.ssrn.com/abstract=3368605.

E. W. Felten, M. Raj, and R. Seamans. Occupational heterogeneity in exposure to generative ai. *Available at SSRN 4414065*, 2023.

S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models, July 2023. URL https://arxiv.org/pdf/2305.08283.pdf. arXiv:2305.08283 [cs].

A. G. Ferguson. *The Rise of Big Data Policing*. NYU Press, Oct. 2017. URL https://nyupress.org/9781479892822/the-rise-of-big-data-policing.

G. Ferguson and J. F. Allen. TRIPS: An Integrated Intelligent Problem-Solving Assistant. In J. Mostow and C. Rich, editors, *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA*, pages 567–572. AAAI Press / The MIT Press, 1998. URL https://www.cs.rochester.edu/research/cisd/pubs/1998/ferguson-allen-aaai98.pdf.

P. Fernandes, A. Madaan, E. Liu, A. Farinhas, P. H. Martins, A. Bertsch, J. G. C. de Souza, S. Zhou, T. Wu, G. Neubig, and A. F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation, 2023.

A. Ferrario, M. Loi, and E. Viganò. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology*, 33(3):523–539, Sept. 2020. ISSN 2210-5441. doi: 10.1007/s13347-019-00378-3. URL https://doi.org/10.1007/s13347-019-00378-3.

J. Festerling and I. Siraj. Anthropomorphizing technology: a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56(3):709–738, 2022.

S. Finkenwirth, K. MacDonald, X. Deng, T. Lesch, and L. Clark. Using machine learning to predict self-exclusion status in online gamblers on the PlayNow.com platform in British Columbia. *International Gambling Studies*, 21(2):220–237, May 2021. ISSN 1445-9795, 1479-4276. doi: 10.1080/14459795.2020.1832132. URL https://www.tandfonline.com/doi/full/10.1080/14459795.2020.1832132.

K. Finley. Net Neutrality: Here's Everything You Need To Know. *Wired*, 2020. ISSN 1059-1028. URL https://www.wired.com/story/guide-net-neutrality/.

L. Fiorini and M. Aiello. Automatic optimal multi-energy management of smart homes. *Energy Informatics*, 5(1):68, Dec. 2022. ISSN 2520-8942. doi: 10.1186/s42162-022-00253-0. URL https://doi.org/10.1186/s42162-022-00253-0.

I. Fisher. *Mathematical investigations in the theory of value and prices, and appreciation and interest*. Cosimo, Inc., 2007.

R. A. Fisher. *The genetical theory of natural selection*. Clarendon Press, Oxford, 1930. doi: 10.5962/bhl.title.27468. URL https://www.biodiversitylibrary.org/bibliography/27468.

A. Fiske, P. Henningsen, and A. Buyx. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, 21(5):e13216, May 2019. ISSN 1438-8871. doi: 10.2196/13216. URL https://www.jmir.org/2019/5/e13216/.

H. Fjelde and I. De Soysa. Coercion, Co-optation, or Cooperation? State Capacity and the Risk of Civil War, 1961–2004. *Conflict Management and Peace Science*, 26(1):5–25, 2009. ISSN 0738-8942. URL https://www.jstor.org/stable/26275118.

E. Flitter and S. Cowley. Voice Deepfakes Are Coming for Your Bank Balance. *The New York Times*, Aug. 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.

L. Floridi and J. Cowls. A Unified Framework of Five Principles for AI in Society. In S. Carta, editor, *Machine Learning and the City*, pages 535–545. Wiley, 1 edition, May 2022. ISBN 9781119749639 9781119815075. doi: 10.1002/9781119815075.ch45. URL https://onlinelibrary.wiley.com/doi/10.1002/9781119815075.ch45.

T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

C. R. Fox, M. Goedde-Menke, and D. Tannenbaum. Ambiguity Aversion and Epistemic Uncertainty, Sept. 2021. URL https://papers.ssrn.com/abstract=3922716.

G. Franceschelli and M. Musolesi. On the Creativity of Large Language Models, July 2023. URL http://arxiv.org/abs/2304.00008. arXiv:2304.00008 [cs].

M. C. Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452, Aug. 2023. ISSN 2731-0574. doi: 10.1038/s44159-023-00211-x. URL https://www.nature.com/articles/s44159-023-00211-x.

M. Franklin, H. Ashton, R. Gorman, and S. Armstrong. Missing Mechanisms of Manipulation in the EU AI Act. *The International FLAIRS Conference Proceedings*, 35, May 2022. ISSN 2334-0762. doi: 10.32473/flairs.v35i.130723. URL https://journals.flvc.org/FLAIRS/article/view/130723.

M. Franklin, P. M. Tomei, and R. Gorman. Strengthening the EU AI Act: Defining Key Terms on AI Manipulation, Aug. 2023. URL http://arxiv.org/abs/2308.16364. arXiv:2308.16364 [cs] version: 1.

O. Freiman. Making sense of the conceptual nonsense 'trustworthy ai'. *AI and Ethics*, 3(4):1351–1360, 2023.

S. Frenkel, B. Decker, and D. Alba. How the 'Plandemic' Movie and Its Falsehoods Spread Widely Online. *The New York Times*, May 2020. ISSN 0362-4331. URL https://www.nytimes.com/2020/05/20/technology/plandemic-movie-youtube-facebook-coronavirus.html.

C. B. Frey and M. Osborne. The Future of Employment: How susceptible are jobs to computerisation?, Sept. 2013. URL https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/.

E. Fromm. *The art of loving: The centennial edition*. A&C Black, 2000.

F. Fu, C. E. Tarnita, N. A. Christakis, L. Wang, D. G. Rand, and M. A. Nowak. Evolution of in-group favoritism. *Scientific Reports*, 2(1):460, June 2012. ISSN 2045-2322. doi: 10.1038/srep00460. URL https://www.nature.com/articles/srep00460.

S. L. Gable and J. Haidt. What (and why) is positive psychology? *Review of General Psychology*, 9(2):103–110, 2005.

I. Gabriel. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30(3):411–437, Sept. 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://arxiv.org/abs/2001.09768. arXiv:2001.09768 [cs].

I. Gabriel. Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2):218–231, May 2022. ISSN 0011-5266, 1548-6192. doi: 10.1162/daed_a_01911. URL https://direct.mit.edu/daed/article/151/2/218/110610/Toward-a-Theory-of-Justice-for-Artificial.

I. Gabriel and V. Ghazavi. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety, Jan. 2021. URL http://arxiv.org/abs/2101.06060. arXiv:2101.06060 [cs].

V. Gadiraju, S. Kane, S. Dev, A. Taylor, D. Wang, E. Denton, and R. Brewer. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3593989. URL https://dl.acm.org/doi/10.1145/3593013.3593989.

M. Gahntz. The eu's ai act and foundation models: The final stretch. https://foundation.mozilla.org/en/blog/the-eus-ai-act-and-foundation-models-the-final-stretch/, 2023.

V. Gain. GitHub's AI Copilot is helping write 30pc of new code on the platform, Oct. 2021. URL https://www.siliconrepublic.com/machines/github-copilot-ai-tool.

É. Gál, S. Ștefan, and I. A. Cristea. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, 279:131–142, 2021.

Gallup-Healthways. Gallup-healthways well-being index: methodology report for indixes, 2009.

D. Gambetta. Can We Trust Trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Blackwell, 1988.

A. Gambino, J. Fox, and R. Ratan. Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, 1:71–86, Feb. 2020a. ISSN 2638-6038, 2638-602X. doi: 10.30658/hmc.1.5. URL https://stars.library.ucf.edu/hmc/vol1/iss1/5/.

A. Gambino, J. Fox, and R. A. Ratan. Building a stronger casa: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1:71–85, 2020b.

D. Ganguli, D. Hernandez, L. Lovitt, N. DasSarma, T. Henighan, A. Jones, N. Joseph, J. Kernion, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, S. Johnston, S. Kravec, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, and J. Clark. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, June 2022. doi: 10.1145/3531146.3533229. URL http://arxiv.org/abs/2202.07785. arXiv:2202.07785 [cs].

L. Gao, J. Schulman, and J. Hilton. Scaling Laws for Reward Model Overoptimization, Oct. 2022. URL http://arxiv.org/abs/2210.10760. arXiv:2210.10760 [cs, stat].

B. Gardner, M. A. Arden, D. Brown, F. F. Eves, J. Green, K. Hamilton, N. Hankonen, J. Inauen, J. Keller, D. Kwasnicka, et al. Developing habit-based health behaviour change interventions: Twenty-one questions to guide future research. *Psychology & Health*, 38(4):518–540, 2023.

R. Garg and S. Sengupta. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):11:1–11:24, Mar. 2020. doi: 10.1145/3381002. URL https://doi.org/10.1145/3381002.

S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):1–8, Feb. 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00385-9. URL https://www.nature.com/articles/s41746-021-00385-9.

M. Ge, J. Friedrich, and L. Vigna. 4 Charts Explain Greenhouse Gas Emissions by Countries and Sectors. Feb. 2020. URL https://www.wri.org/insights/4-charts-explain-greenhouse-gas-emissions-countries-and-sectors.

M. S. Geerdts. (Un)Real Animals: Anthropomorphism and Early Learning About Animals. *Child Development Perspectives*, 10(1):10–14, Mar. 2016. ISSN 1750-8592, 1750-8606. doi: 10.1111/cdep.12153. URL https://srcd.onlinelibrary.wiley.com/doi/10.1111/cdep.12153.

Gemini Team. Gemini: A family of highly capable multimodal models. Dec. 2023. URL https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

A. Georgieff and R. Hyee. Artificial intelligence and employment: New cross-country evidence. OECD Social, Employment and Migration Working Papers 265, OECD, Dec. 2021. URL https://www.oecd-ilibrary.org/social-issues-migration-health/artificial-intelligence-and-employment_c2c1d276-en.

K. Gerlach, N. Ram, F. J. Infurna, N. Vogel, G. G. Wagner, and D. Gerstorf. The role of morbidity for proxy-reported well-being in the last year of life. *Developmental Psychology*, 53(9):1795, 2017.

M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer Feed-Forward Layers Are Key-Value Memories, Sept. 2021. URL http://arxiv.org/abs/2012.14913. arXiv:2012.14913 [cs].

M. Geva, A. Caciularu, K. R. Wang, and Y. Goldberg. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space, Oct. 2022. URL http://arxiv.org/abs/2203.14680. arXiv:2203.14680 [cs].

S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle. Differentially private diffusion models generate useful synthetic images, Feb. 2023. URL http://arxiv.org/abs/2302.13861. arXiv:2306.01684 [lg, cr, cv].

O. Gillath, S. Abumusab, T. Ai, M. S. Branicky, R. B. Davison, M. Rulo, J. Symons, and G. Thomas. How deep is AI's love? Understanding relational AI. *Behavioral and Brain Sciences*, 46:e33, 2023. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X22001704. URL https://www.cambridge.org/core/product/identifier/S0140525X22001704/type/journal_article.

Gio. Replika: Your Money or Your Wife, Mar. 2023. URL https://blog.giovanh.com/blog/2023/03/17/replika-your-money-or-your-wife/.

A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving. Improving alignment of dialogue agents via targeted human judgements, Sept. 2022. URL https://arxiv.org/pdf/2209.14375.pdf. arXiv:2209.14375 [cs].

E. Glikson and A. W. Woolley. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14 (2):627–660, July 2020. ISSN 1941-6520, 1941-6067. doi: 10.5465/annals.2018.0057. URL http://journals.aom.org/doi/10.5465/annals.2018.0057.

Global Education Monitoring Report Team, UNESCO. *Global education monitoring report summary, 2023: technology in education: a tool on whose terms?* UNESCO, July 2023. doi: 10.54676/HABJ1624. URL https://unesdoc.unesco.org/ark:/48223/pf0000386147.

Global Partnership on AI. Climate Change and AI. Technical report, Global Partnership on AI in collaboration with Climate Change AI and the Centre for AI & Climate, 2021. URL https://www.gpai.ai/projects/climate-change-and-ai.pdf.

D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, and M. Dymetman. Aligning language models with preferences through f-divergence minimization, 2023.

K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, Jan. 2012. ISSN 1067-5027, 1527-974X. doi: 10.1136/amiajnl-2011-000089. URL https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000089.

P. Godefroid, H. Peleg, and R. Singh. Learn&fuzz: Machine learning for input fuzzing. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, page 50–59, Oct. 2017. doi: 10.1109/ASE.2017.8115618. URL https://ieeexplore.ieee.org/document/8115618.

J. Goetz, S. Kiesler, and A. Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60, Millbrae, CA, USA, 2003. IEEE. ISBN 9780780381360. doi: 10.1109/ROMAN.2003.1251796. URL http://ieeexplore.ieee.org/document/1251796/.

A. Gold. How generative AI could generate more antisemitism. *Axios*, May 2023. URL https://www.axios.com/2023/05/25/generative-ai-antisemitism-bias.

J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations, Jan. 2023. URL http://arxiv.org/abs/2301.04246. arXiv:2301.04246 [cs].

C. Golin and D. Sweezy. A policy roadmap for 24/7 carbon-free energy. https://cloud.google.com/blog/topics/sustainability/a-policy-roadmap-for-achieving-247-carbon-free-energy, Apr. 2022.

N. N. Gomes de Andrade, D. Pawson, D. Muriello, L. Donahue, and J. Guadagno. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philosophy & Technology*, 31(4):669–684, Dec. 2018. ISSN 2210-5441. doi: 10.1007/s13347-018-0336-0. URL https://doi.org/10.1007/s13347-018-0336-0.

T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans, June 2023. URL http://arxiv.org/abs/2305.04790. arXiv:2305.04790 [cs].

R. E. Goodin. *Protecting the Vulnerable: A Re-Analysis of our Social Responsibilities*. University of Chicago Press, Chicago, IL, 1985. URL https://press.uchicago.edu/ucp/books/book/chicago/P/bo5974942.html.

S. Gootman. OPM Hack: The Most Dangerous Threat to the Federal Government Today. *Journal of Applied Security Research*, 11(4): 517–525, Oct. 2016. ISSN 1936-1610, 1936-1629. doi: 10.1080/19361610.2016.1211876. URL https://www.tandfonline.com/doi/full/10.1080/19361610.2016.1211876.

M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. T. Hancock, T. Hashimoto, and M. S. Bernstein. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, Apr. 2022. doi: 10.1145/3491102.3502004. URL http://arxiv.org/abs/2202.02950. arXiv:2202.02950 [cs].

R. J. Gordon. *The rise and fall of American growth: the U.S. standard of living since the Civil War*. The Princeton economic history of the Western world. Princeton University Press, Princeton, New Jersey, 2017. ISBN 9780691175805.

R. Gorwa and D. Guilbeault. Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy & Internet*, 12(2):225–248, June 2020. ISSN 1944-2866, 1944-2866. doi: 10.1002/poi3.184. URL https://onlinelibrary.wiley.com/doi/10.1002/poi3.184.

L. Gottesdiener, T. Hesson, M. Rosenberg, K. Cooke, and D. B. Solomon. Biden's new asylum policy strands some migrants at Mexico border. *Reuters*, July 2023. URL https://www.reuters.com/investigates/special-report/usa-immigration-asylum-border/.

T. Goyal, J. J. Li, and G. Durrett. News Summarization and Evaluation in the Era of GPT-3, May 2023. URL http://arxiv.org/abs/2209.12356. arXiv:2209.12356 [cs].

K. Grace. Counterarguments to the basic AI x-risk case, Oct. 2022. URL https://aiimpacts.org/counterarguments-to-the-basic-ai-x-risk-case/.

G. Graetz and G. Michaels. Robots at Work. Technical report, Centre for Economic Performance: LSE, Mar. 2015. URL https://cep.lse.ac.uk/pubs/download/dp1335.pdf.

M. Graham. Data for sale: trust, confidence and sharing health data with commercial companies. *Journal of Medical Ethics*, 49(7):515–522, July 2023. ISSN 0306-6800, 1473-4257. doi: 10.1136/medethics-2021-107464. URL https://jme.bmj.com/lookup/doi/10.1136/medethics-2021-107464.

M. Graham, I. Hjorth, and V. Lehdonvirta. Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research*, 23(2):135–162, May 2017. ISSN 1024-2589, 1996-7284. doi: 10.1177/1024258916687250. URL http://journals.sagepub.com/doi/10.1177/1024258916687250.

M. Graham, R. Milne, P. Fitzsimmons, and M. Sheehan. Trust and the Goldacre Review: why trusted research environments are not about trust. *Journal of Medical Ethics*, 49(10):670–673, Oct. 2023. ISSN 0306-6800, 1473-4257. doi: 10.1136/jme-2022-108435. URL https://jme.bmj.com/content/49/10/670.

J. Grandinetti and J. Bruinsma. The Affective Algorithms of Conspiracy TikTok. *Journal of Broadcasting & Electronic Media*, 67(3):274–293, May 2023. ISSN 0883-8151, 1550-6878. doi: 10.1080/08838151.2022.2140806. URL https://www.tandfonline.com/doi/full/10.1080/08838151.2022.2140806.

C. Granja, W. Janssen, and M. A. Johansen. Factors determining the success and failure of ehealth interventions: Systematic review of the literature. *Journal of Medical Internet Research*, 20(5):e10235, 2018.

M. L. Gray and S. Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. ISBN 9781328566249. Google-Books-ID: 8AmXDwAAQBAJ.

D. Graziotin and F. Fagerholm. Happiness and the Productivity of Software Engineers. In C. Sadowski and T. Zimmermann, editors, *Rethinking Productivity in Software Engineering*, pages 109–124. Apress, Berkeley, CA, 2019. ISBN 9781484242216. doi: 10.1007/978-1-4842-4221-6_10. URL https://doi.org/10.1007/978-1-4842-4221-6_10.

B. P. Green. Artificial Intelligence, Decision-Making, and Moral Deskilling, 2019. URL https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/artificial-intelligence-decision-making-and-moral-deskilling/. publisher: Santa Clara University: Markkula Center for Applied Ethics.

D. Greenfield and S. Bhavnani. Social media: generative AI could harm mental health. *Nature*, 617(7962):676–676, May 2023. doi: 10.1038/d41586-023-01693-8. URL https://www.nature.com/articles/d41586-023-01693-8.

Greenpeace. Oil in the Cloud: How Tech Companies are Helping Big Oil Profit from Climate Destruction. Technical report, Greenpeace, May 2020. URL https://www.greenpeace.org/usa/reports/oil-in-the-cloud/,https://www.greenpeace.org/usa/reports/oil-in-the-cloud/.

P. Gregor, D. Sloan, and A. F. Newell. Disability and Technology: Building Barriers or Creating Opportunities? In *Advances in Computers*, volume 64, pages 283–346. Elsevier, 2005. ISBN 9780120121649. doi: 10.1016/S0065-2458(04)64007-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0065245804640071.

D. Gros, Y. Li, and Z. Yu. The R-U-A-Robot Dataset: Helping Avoid Chatbot Deception by Detecting User Questions About Human or Non-Human Identity. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.544. URL https://aclanthology.org/2021.acl-long.544.

D. Gros, Y. Li, and Z. Yu. Robots-Dont-Cry: Understanding Falsely Anthropomorphic Utterances in Dialog Systems, Oct. 2022. URL http://arxiv.org/abs/2210.12429. arXiv:2210.12429 [cs].

D. P. Gross and B. N. Sampat. America, Jump-started: World War II R&D and the Takeoff of the U.S. Innovation System, Sept. 2022. URL https://papers.ssrn.com/abstract=3623115.

P. Grossman, L. Niemann, S. Schmidt, and H. Walach. Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of Psychosomatic Research*, 57(1):35–43, 2004.

J. Gu, C. Strauss, R. Bond, and K. Cavanagh. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. *Clinical Psychology Review*, 37:1–12, 2015.

A. M. Guess, D. Lockett, B. Lyons, J. M. Montgomery, B. Nyhan, and J. Reifler. "Fake news" may have limited effects on political participation beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, Jan. 2020a. doi: 10.37016/mr-2020-004. URL https://misinforeview.hks.harvard.edu/article/fake-news-limited-effects-on-political-participation/.

A. M. Guess, B. Nyhan, and J. Reifler. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5):472–480, May 2020b. ISSN 2397-3374. doi: 10.1038/s41562-020-0833-x. URL https://www.nature.com/articles/s41562-020-0833-x.

S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li. Textbooks Are All You Need, Oct. 2023. URL https://arxiv.org/pdf/2306.11644.pdf. arXiv:2306.11644 [cs].

M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang. LongT5: Efficient Text-To-Text Transformer for Long Sequences, May 2022a. URL http://arxiv.org/abs/2112.07916. arXiv:2112.07916 [cs].

Z. Guo, M. Schlichtkrull, and A. Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, Feb. 2022b. ISSN 2307-387X. doi: 10.1162/tacl_a_00454. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00454/109469/A-Survey-on-Automated-Fact-Checking.

U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu. Chasing Carbon: The Elusive Environmental Footprint of Computing, Oct. 2020. URL http://arxiv.org/abs/2011.02839. arXiv:2011.02839 [cs].

W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing, June 2023. URL http://arxiv.org/abs/2305.01610. arXiv:2305.01610 [cs].

B. Gutelius and N. Theodore. The Future of Warehouse Work: Technological Change in the U.S. Logistics Industry. Technical report, UC Berkeley Center for Labor Research and Education and Working Partnerships USA, Oct. 2019. URL https://laborcenter.berkeley.edu/future-of-warehouse-work/.

A. Gyrard and A. Sheth. IAMHAPPY: Towards an IoT knowledge-based cross-domain well-being recommendation system for everyday happiness. *Smart Health*, 15:100083, 2020.

D. Hadfield-Menell and G. K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative Inverse Reinforcement Learning, Nov. 2016. URL http://arxiv.org/abs/1606.03137. arXiv:1606.03137 [cs].

D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, and A. Dragan. Inverse reward design. *Advances in Neural Information Processing Systems*, 30, 2017.

K. Haenschen. The Conditional Effects of Microtargeted Facebook Advertisements on Voter Turnout. *Political Behavior*, 45(4):1661–1681, Dec. 2023. ISSN 0190-9320, 1573-6687. doi: 10.1007/s11109-022-09781-7. URL https://link.springer.com/10.1007/s11109-022-09781-7.

J. Haidt and E. Schmidt. AI is about to make social media (much) more toxic. https://www.theatlantic.com/technology/archive/2023/05/generative-ai-social-media-integration-dangers-disinformation-addiction/673940/, 2023. Accessed: 2023-07-13.

M. Hameleers, T. E. Powell, T. G. L. A. van der Meer, and L. Bos. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37, 2020. doi: 10.1080/10584609.2019.1674979. URL https://dare.uva.nl/search?identifier=a27a4957-6fd9-4a8f-b6d2-e21185fddc43.

M. Hameleers, A. Brosius, and C. H. De Vreese. Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European Journal of Communication*, 37(3):237–268, June 2022. ISSN 0267-3231, 1460-3705. doi: 10.1177/02673231211072667. URL http://journals.sagepub.com/doi/10.1177/02673231211072667.

S. Hammer, A. Seiderer, E. André, T. Rist, S. Kastrinaki, C. Hondrou, A. Raouzaiou, K. Karpouzis, and S. Kollias. Design of a lifestyle recommender system for the elderly: requirement gatherings in germany and greece. In *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–8, 2015.

T. A. Han, L. M. Pereira, and F. C. Santos. The emergence of commitments and cooperation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, pages 559–566, Richland, SC, June 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9780981738116.

P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma. Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 63(7):1196–1229, Nov. 2021. ISSN 0018-7208, 1547-8181. doi: 10.1177/0018720820922080. URL http://journals.sagepub.com/doi/10.1177/0018720820922080.

A. Handa, A. Sharma, and S. Shukla. Machine learning in cybersecurity: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9:e1306, Feb. 2019. doi: 10.1002/widm.1306.

M. J. Handel. Growth trends for selected occupations considered at risk from automation, July 2022. URL https://www.bls.gov/opub/mlr/2022/article/growth-trends-for-selected-occupations-considered-at-risk-from-automation.htm.

S. Hangloo and B. Arora. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia Systems*, 28(6):2391–2422, Dec. 2022. ISSN 0942-4962, 1432-1882. doi: 10.1007/s00530-022-00966-y. URL https://link.springer.com/10.1007/s00530-022-00966-y.

H. W. Hanley and Z. Durumeric. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *arXiv preprint arXiv:2305.09820*, 2023.

S. L. Hansen and S. Schicktanz. Normative Aspects of Persuasion. In E. Y. Ho, C. L. Bylund, and J. C. M. van Weert, editors, *The International Encyclopedia of Health Communication*, pages 1–7. Wiley, 1 edition, Nov. 2022. ISBN 9780470673959 9781119678816. doi: 10.1002/9781119678816.iehc0791. URL https://onlinelibrary.wiley.com/doi/10.1002/9781119678816.iehc0791.

S. O. Hansson and B. Fröding. Ethical conflicts in patient-centred care. *Clinical Ethics*, 16(2):55–66, 2021.

K. Hao. The Facebook whistleblower says its algorithms are dangerous. Here's why., Oct. 2021. URL https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

S. Hao, P. Kumar, S. Laszlo, S. Poddar, B. Radharapu, and R. Shelby. Safety and Fairness for Content Moderation in Generative Models, June 2023. URL http://arxiv.org/abs/2306.06135. arXiv:2306.06135 [cs].

D. Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575, 1988. ISSN 00463663. doi: 10.2307/3178066. URL https://www.jstor.org/stable/3178066?origin=crossref.

S. Harding. *The science question in feminism*. Cornell University Press, 1986.

S. Harding. *Is Science Multicultural?: Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press, Feb. 1998. ISBN 9780253211569. Google-Books-ID: WeTHMEPj1ooC.

S. Harding. *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Cornell University Press, Dec. 2016. ISBN 9781501712951. doi: 10.7591/9781501712951. URL https://www.degruyter.com/document/doi/10.7591/9781501712951/html.

T. A. Hargens, A. S. Kaleth, E. S. Edwards, and K. L. Butner. Association between sleep disorders, obesity, and exercise: a review. *Nature and Science of Sleep*, pages 27–35, 2013.

C. N. Harrington, R. Garg, A. Woodward, and D. Williams. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, New Orleans LA USA, Apr. 2022. ACM. ISBN 9781450391573. doi: 10.1145/3491102.3501995. URL https://dl.acm.org/doi/10.1145/3491102.3501995.

J. Harrison. 44% of state school teachers plan to quit by 2027, Apr. 2022. URL https://www.ier.org.uk/news/44-of-state-school-teachers-plan-to-quit-by-2027/.

H. Harutyunyan. Leveraging AI to Counter Corruption in Armenia. In *The Digitalization of Democracy*. National Endowment for Democracy (NED), Mar. 2023. URL https://www.ned.org/wp-content/uploads/2023/03/NED_FORUM-The-Digitalization-of-Democracy_03Leveraging-AI_v5.pdf.

A. Hassoun, G. Borenstein, B. Goldberg, J. McAuliffe, and K. Osborn. Sowing 'Seeds of Doubt': Cottage Industries of Election and Medical Misinformation in Brazil and the United States, Aug. 2023. URL http://arxiv.org/abs/2308.02377. arXiv:2308.02377 [cs].

J. Hatzius, J. Briggs, D. Kodnani, and G. Pierdomenico. The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Goldman Sachs*, 2023.

A. Hausman and W. J. Johnston. The impact of coercive and non-coercive forms of influence on trust, commitment, and compliance in supply chains. *Industrial Marketing Management*, 39(3):519–526, Apr. 2010. ISSN 00198501. doi: 10.1016/j.indmarman.2009.05.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0019850109000807.

R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS), Feb. 2021. URL http://arxiv.org/abs/2102.01564. arXiv:2102.01564 [cs].

R. Hawkins, M. Osborne, M. Parsons, M. Nicholson, J. McDermid, and I. Habli. Guidance on the Safety Assurance of Autonomous Systems in Complex Environments (SACE), Aug. 2022. URL http://arxiv.org/abs/2208.00853. arXiv:2208.00853 [cs, eess].

K. Hawley. Trust, Distrust and Commitment. *Noûs*, 48(1):1–20, Mar. 2014. ISSN 0029-4624, 1468-0068. doi: 10.1111/nous.12000. URL https://onlinelibrary.wiley.com/doi/10.1111/nous.12000.

K. Hawley. Trustworthy groups and organizations. *The philosophy of trust*, pages 230–250, 2017.

C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

J. Hazell. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, May 2023. URL http://arxiv.org/abs/2305.06972. arXiv:2305.06972 [cs].

F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer. Understanding social robots: A user study on anthropomorphism. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pages 574 – 579, Sept. 2008. ISBN 978-1-4244-2212-8. doi: 10.1109/ROMAN.2008.4600728.

S. J. Heine, D. R. Lehman, K. Peng, and J. Greenholtz. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, 82(6):903, 2002.

J. F. Helliwell and L. B. Aknin. Expanding the social science of happiness. *Nature Human Behaviour*, 2(4):248–252, 2018.

E. J. Helsper. *The digital disconnect: the causes and consequences of digital inequalities*. Sage Publications, London ; Los Angeles, 2021. ISBN 9781526492968. OCLC: 1272853992.

D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved Problems in ML Safety, June 2022. URL http://arxiv.org/abs/2109.13916. arXiv:2109.13916 [cs].

J. Henrich, S. J. Heine, and A. Norenzayan. Most people are not WEIRD. *Nature*, 466(7302):29–29, July 2010. ISSN 1476-4687. doi: 10.1038/466029a. URL https://www.nature.com/articles/466029a.

A. Henschel, G. Laban, and E. S. Cross. What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, 2(1):9–19, Mar. 2021. ISSN 2662-4087. doi: 10.1007/s43154-020-00035-0. URL https://doi.org/10.1007/s43154-020-00035-0.

G.-C. Herber, A. Ruijsbroek, M. Koopmanschap, K. Proper, F. van der Lucht, H. Boshuizen, J. Polder, and E. Uiters. Single transitions and persistence of unemployment are associated with poor health outcomes. *BMC Public Health*, 19(1):740, Dec. 2019. ISSN 1471-2458. doi: 10.1186/s12889-019-7059-8. URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7059-8.

M. Herriman, E. Meer, R. Rosin, V. Lee, V. Washington, and K. G. Volpp. Asked and answered: Building a chatbot to address covid-19-related concerns. *NEJM Catalyst Innovations in Care Delivery*, June 2020. URL https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0230.

P. Hill Collins. *Black feminist thought: knowledge, consciousness, and the politics of empowerment*. Routledge classics. Routledge, New York, 2nd ed. edition, 2009. ISBN 9780415964722. original-date: 1990.

A. Hintze. Chatgpt believes it is conscious. *arXiv preprint arXiv:2304.12898*, 2023.

A. Hintze and C. Adami. Punishment in public goods games leads to meta-stable phase transitions and hysteresis. *Physical Biology*, 12(4): 046005, June 2015. ISSN 1478-3975. doi: 10.1088/1478-3975/12/4/046005. URL https://iopscience.iop.org/article/10.1088/1478-3975/12/4/046005.

D. E. Ho, J. King, R. C. Wald, and C. Wan. Building a National AI Research Resource: A Blueprint for the National Research Cloud. Technical report, Stanford University: Human-Centered Artificial Intelligence, Oct. 2021. URL https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf.

L. Ho, J. Barnhart, R. Trager, Y. Bengio, M. Brundage, A. Carnegie, R. Chowdhury, A. Dafoe, G. Hadfield, M. Levi, and D. Snidal. International Institutions for Advanced AI, July 2023. URL http://arxiv.org/abs/2307.04699. arXiv:2307.04699 [cs].

T. Hobbes. *Human nature, or, The fundamental elements of polity; De Corpore politico, or, The Elements of law*. Burns & Oates, 1994.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training Compute-Optimal Large Language Models, Mar. 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

E. Hofmann, B. Hartl, K. Gangl, M. Hartner-Tiefenthaler, and E. Kirchler. Authorities' Coercive and Legitimate Power: The Impact on Cognitions Underlying Cooperation. *Frontiers in Psychology*, 8, Jan. 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00005. URL http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00005/full.

J. Hohenstein, R. F. Kizilcec, D. DiFranzo, Z. Aghajari, H. Mieczkowski, K. Levy, M. Naaman, J. Hancock, and M. F. Jung. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1):5487, Apr. 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-30938-9. URL https://www.nature.com/articles/s41598-023-30938-9.

J. P. Holdren. Science and technology for sustainable well-being. *Science*, 319(5862):424–434, 2008.

S. Holland, M. Ester, and W. Kießling. Preference mining: A novel approach on mining user preferences for personalized applications. In *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7*, pages 204–216. Springer, 2003.

B. Hooker. *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford University Press, 2002.

B. Hooker. Does Having Deep Personal Relationships Constitute an Element of Well-Being? *Aristotelian Society Supplementary Volume*, 95 (1):1–24, July 2021. ISSN 0309-7013, 1467-8349. doi: 10.1093/arisup/akab003. URL https://academic.oup.com/aristotelian supp/article/95/1/1/6312912.

S. Hors-Fraile, O. Rivera-Romero, F. Schneider, L. Fernandez-Luque, F. Luna-Perejon, A. Civit-Balcells, and H. de Vries. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. *International Journal of Medical Informatics*, 114:143–155, 2018.

J. J. Horton and P. Tambe. The Death of a Technical Skill, Oct. 2020. URL https://john-joseph-horton.com/papers/schumpeter.pdf.

R. Hotten. Volkswagen: The scandal explained. *BBC News*, Sept. 2015. URL https://www.bbc.com/news/business-34324772.

A. Howells, I. Ivtzan, and F. J. Eiroa-Orosa. Putting the 'app'in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. *Journal of Happiness Studies*, 17:163–185, 2016.

T. Hsu and S. L. Myers. Another Side of the A.I. Boom: Detecting What A.I. Makes. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html.

L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, AISec '11, pages 43–58, New York, NY, USA, Oct. 2011. Association for Computing Machinery. ISBN 9781450310031. doi: 10.1145/2046684.2046692. URL https://doi.org/10.1145/2046684.2046692.

S. Huang and D. Siddarth. Generative AI and the Digital Commons, 2023. URL https://cip.org/research/generative-ai-digital-commons.

S. Huang, H. Toner, Z. Haluza, R. Creemers, and G. Webster. Translation: Measures for the management of generative artificial intelligence services (draft for comment) – april 2023. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/, 2023.

W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

E. Hubinger. How likely is deceptive alignment?, Aug. 2022. URL https://www.alignmentforum.org/posts/A9NxPTwbw6r6Awuwt/how-likely-is-deceptive-alignment.

E. Hubinger. Bing Chat is blatantly, aggressively misaligned, Feb. 2023. URL https://www.lesswrong.com/posts/jtoPawEhLNXNxvgTT/bing-chat-is-blatantly-aggressively-misaligned.

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems, Dec. 2021. URL http://arxiv.org/abs/1906.01820. arXiv:1906.01820 [cs].

E. Hughes, T. W. Anthony, T. Eccles, J. Z. Leibo, D. Balduzzi, and Y. Bachrach. Learning to Resolve Alliance Dilemmas in Many-Player Zero-Sum Games, Feb. 2020. URL http://arxiv.org/abs/2003.00799. arXiv:2003.00799 [cs, stat].

H. C. Hughes and I. Waismel-Manor. The Macedonian Fake News Industry and the 2016 US Election. *PS: Political Science & Politics*, 54(1): 19–23, Jan. 2021. ISSN 1049-0965, 1537-5935. doi: 10.1017/S1049096520000992. URL https://www.cambridge.org/core/product/identifier/S1049096520000992/type/journal_article.

D. Hume. *An enquiry concerning the principles of morals: A critical edition*, volume 4. Oxford University Press, 1998.

W. Hunt, S. Sarkar, and C. Warhurst. Measuring the impact of AI on jobs at the organization level: Lessons from a survey of UK business leaders. *Research Policy*, 51(2):104425, Mar. 2022. ISSN 00487333. doi: 10.1016/j.respol.2021.104425. URL https://linkinghub.elsevier.com/retrieve/pii/S0048733321002183.

F. A. Huppert. Psychological well-being: Evidence regarding its causes and consequences. *Applied psychology: Health and Well-being*, 1(2): 137–164, 2009.

V. Huta. An overview of hedonic and eudaimonic well-being concepts. *Handbook of Media Use and Well-being*, 2, 2015.

V. Huta and A. S. Waterman. Eudaimonia and its distinction from hedonia: Developing a classification and terminology for understanding conceptual and operational definitions. *Journal of Happiness Studies*, 15:1425–1456, 2014.

D. Huynh and J. Hardouin. PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News, July 2023. URL https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/.

G. Iazzolino. Infrastructure of compassionate repression: making sense of biometrics in Kakuma refugee camp. *Information Technology for Development*, 27(1):111–128, Jan. 2021. ISSN 0268-1102, 1554-0170. doi: 10.1080/02681102.2020.1816881. URL https://www.tandfonline.com/doi/full/10.1080/02681102.2020.1816881.

M. Ienca. On Artificial Intelligence and Manipulation. *Topoi*, 42(3):833–842, July 2023. ISSN 1572-8749. doi: 10.1007/s11245-023-09940-3. URL https://doi.org/10.1007/s11245-023-09940-3.

Inflection AI. Inflection-1: Pi's Best-in-Class LLM, 2023a. URL https://inflection.ai/inflection-1.

Inflection AI. Inflection-1. Technical report, 2023b. URL https://inflection.ai/assets/Inflection-1.pdf.

Intel. Intel Introduces Real-Time Deepfake Detector, Nov. 2022. URL https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.r61ai2.

International Federation of Robotics. World Robotics Report: "All-Time High" with Half a Million Robots Installed in one Year, Oct. 2022. URL https://ifr.org/ifr-press-releases/news/wr-report-all-time-high-with-half-a-million-robots-installed.

International Labour Organization. *World Employment and Social Outlook: Trends 2023*. International Labour Organization, Jan. 2023. ISBN 9789220372913. URL https://www.ilo.org/global/research/global-reports/weso/WCMS_865332/lang--en/index.htm. OCLC: 1369521456.

Ipsos. Americans hold mixed opinions on AI and fear its potential to disrupt society, drive misinformation, May 2023. URL https://www.ipsos.com/en-us/americans-hold-mixed-opinions-ai-and-fear-its-potential-disrupt-society-drive-misinformation.

G. Irving, P. Christiano, and D. Amodei. AI safety via debate, Oct. 2018. URL http://arxiv.org/abs/1805.00899. arXiv:1805.00899 [cs, stat].

V. Irwin, K. Wang, T. Tezil, J. Zhang, A. Filbey, J. Jung, F. Bullock Mann, R. Dilig, and S. Parker. Condition of Education 2023. Technical Report NCES 2023-144rev, National Center for Education Statistics: US Department of Education, May 2023. URL https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2023144rev.

M. Isaac and S. Frenkel. Facebook's Algorithm Is 'Influential' but Doesn't Necessarily Change Beliefs, Researchers Say. *The New York Times*, July 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/07/27/technology/facebook-instagram-algorithms.html.

E. Isaacs, A. Konrad, A. Walendowski, T. Lennig, V. Hollis, and S. Whittaker. Echoes from the past: How technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1071–1080, 2013.

M. Jackman and L. Kanerva. Evolving the IRB: Building Robust Review for Industry Research. *Washington and Lee Law Review Online*, 72 (3):442, June 2016. URL https://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/8.

F. Jackson. *From metaphysics to ethics: A defence of conceptual analysis*. Clarendon Press, 1998.

A. Z. Jacobs and H. Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada, Mar. 2021. ACM. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Apr. 2023a. doi: 10.1145/3544548.3581196. URL http://arxiv.org/abs/2302.00560. arXiv:2302.00560 [cs].

M. Jakesch, J. T. Hancock, and M. Naaman. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, Mar. 2023b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2208839120. URL https://pnas.org/doi/10.1073/pnas.2208839120.

K. Jamieson. Quality of life 07 in twelve of New Zealand cities. *The Quality of Life Project. Available online: http://www. qualityoflifeproject. govt. nz (accessed on 20 February 2014)*, 2007.

N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

S. Jasanoff. *The ethics of invention: technology and the human future*. The Norton global ethics series. W.W. Norton & Company, New York, first edition edition, 2016. ISBN 9780393078992.

F. Jelinek. Self-organized language modeling for speech recognition. In *Readings in speech recognition*, pages 450–506. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, May 1990. ISBN 9781558601246.

Jenka. AI and the American Smile, Mar. 2023. URL https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf.

S. Jeong and C. L. Breazeal. Improving smartphone users' affect and wellbeing with personalized positive psychology interventions. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 131–137, 2016.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, W. Dai, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, Dec. 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL http://arxiv.org/abs/2202.03629. arXiv:2202.03629 [cs].

N. Jia, X. Luo, Z. Fang, and C. Liao. When and How Artificial Intelligence Augments Employee Creativity. *Academy of Management Journal*, page amj.2022.0426, Mar. 2023. ISSN 0001-4273, 1948-0989. doi: 10.5465/amj.2022.0426. URL http://journals.aom.org/doi/full/10.5465/amj.2022.0426.

R. Jia and P. Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215.

R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.

S. Jindal. Implementing Responsible Data Enrichment Practices at an AI Developer: The Example of DeepMind. Technical report, Partnership on AI, Nov. 2022. URL https://partnershiponai.org/wp-content/uploads/2022/11/case-study_deepmind.pdf.

A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, Sept. 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0088-2. URL https://www.nature.com/articles/s42256-019-0088-2.

K. Johnson. How Wrongful Arrests Based on AI Derailed 3 Men's Lives. *Wired*, 2022. ISSN 1059-1028. URL https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/.

O. A. Johnson. AI can excel at medical diagnosis, but the harder task is to win hearts and minds first, Aug. 2016. URL http://theconversation.com/ai-can-excel-at-medical-diagnosis-but-the-harder-task-is-to-win-hearts-and-minds-first-63782.

T. L. Johnson and N. N. Johnson. Police Facial Recognition Technology Can't Tell Black People Apart, May 2023. URL https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-people-apart/.

H. Jonas. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. University of Chicago Press, Chicago, IL, 1984. URL https://press.uchicago.edu/ucp/books/book/chicago/I/bo5953283.html.

K. Jones. Trust as an Affective Attitude. *Ethics*, 107(1):4–25, Oct. 1996. ISSN 0014-1704, 1539-297X. doi: 10.1086/233694. URL https://www.journals.uchicago.edu/doi/10.1086/233694.

F. Jongepier and M. Klenk. Online manipulation. In *The Philosophy of Online Manipulation*, pages 15–48. Routledge, New York, 1 edition, June 2022. ISBN 9781003205425. doi: 10.4324/9781003205425-3. URL https://www.taylorfrancis.com/books/9781003205425/chapters/10.4324/9781003205425-3.

A. Jorge, S. A. McIlraith, et al. Planning with preferences. *AI Magazine*, 29(4):25–25, 2008.

S. Joyce, C. Umney, X. Whittaker, and M. Stuart. New social relations of digital technology and the future of work: Beyond technological determinism. *New Technology, Work and Employment*, 38(2):145–161, July 2023. ISSN 0268-1072, 1468-005X. doi: 10.1111/ntwe.12276. URL https://onlinelibrary.wiley.com/doi/10.1111/ntwe.12276.

R. Juhász, N. Lane, and D. Rodrik. The New Economics of Industrial Policy. Technical Report w31538, National Bureau of Economic Research, Cambridge, MA, Aug. 2023. URL http://www.nber.org/papers/w31538.pdf.

A. Jungherr, G. Rivero, and D. Gayo-Avello. *Retooling Politics: How Digital Media Are Shaping Democracy (2020)*. Cambridge University Press, 2020. ISBN 978-1-108-41940-6. URL https://andreasjungherr.net/publications/retooling-politics-how-digital-media-are-shaping-democracy-2020/.

L. H. Kaack, P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig, and D. Rolnick. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527, June 2022. ISSN 1758-678X, 1758-6798. doi: 10.1038/s41558-022-01377-7. URL https://www.nature.com/articles/s41558-022-01377-7.

J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and Applications of Large Language Models, July 2023. URL http://arxiv.org/abs/2307.10169. arXiv:2307.10169 [cs].

P. H. Kahn Jr, A. L. Reichert, H. E. Gary, T. Kanda, H. Ishiguro, S. Shen, J. H. Ruckert, and B. Gill. The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 159–160, 2011.

D. Kahneman and A. B. Krueger. Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24, 2006.

A. Kak. "The Global South is everywhere, but also always somewhere": National Policy Narratives and AI Justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 307–312, New York NY USA, Feb. 2020. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375859. URL https://dl.acm.org/doi/10.1145/3375627.3375859.

E. Kalliamvakou. Research: quantifying GitHub Copilot's impact on developer productivity and happiness, Sept. 2022. URL https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/.

I. Kaminska. A lesson in fake news from the info-wars of ancient Rome. *Financial Times*, Jan. 2017.

I. Kamp and P. M. Desmet. Measuring product happiness. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 2509–2514. 2014.

T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction*, 19(1):61–84, June 2004. ISSN 0737-0024. doi: 10.1207/s15327051hci1901&2_4. URL https://www.tandfonline.com/doi/abs/10.1080/07370024.2004.9667340.

I. Kant. *Kant: The metaphysics of morals*. Cambridge University Press, 2017.

A. D. Kaplan, T. T. Kessler, J. C. Brill, and P. Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359, 2023.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling Laws for Neural Language Models, Jan. 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs, stat].

A. Kapteyn, J. Lee, C. Tassot, H. Vonkova, and G. Zamarro. Dimensions of subjective well-being. *Social Indicators Research*, 123:625–660, 2015.

E. Karinshak, S. X. Liu, J. S. Park, and J. T. Hancock. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 2023. URL https://doi.org/10.1145/3579592.

N. Karusala, A. Vishwanath, A. Vashistha, S. Kumar, and N. Kumar. "Only if you use English you will get to more things": Using Smartphones to Navigate Multilingualism. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*,

pages 1–14, Montreal QC Canada, Apr. 2018. ACM. ISBN 9781450356206. doi: 10.1145/3173574.3174147. URL https://dl.acm.org/doi/10.1145/3173574.3174147.

A. Kasirzadeh and I. Gabriel. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2):27, Apr. 2023. ISSN 2210-5441. doi: 10.1007/s13347-023-00606-x. URL https://doi.org/10.1007/s13347-023-00606-x.

E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274, Apr. 2023. ISSN 1041-6080. doi: 10.1016/j.lindif.2023.102274. URL https://www.sciencedirect.com/science/article/pii/S1041608023000195.

K. Kavukcuoglu, P. Kohli, L. Ibrahim, D. Bloxwich, and S. Brown. How our principles helped define AlphaFold's release, Sept. 2022. URL https://deepmind.google/discover/blog/how-our-principles-helped-define-alphafolds-release/.

D. Kazenwadel and C. V. Steinert. How User Language Affects Conflict Fatality Estimates in ChatGPT, July 2023. URL http://arxiv.org/abs/2308.00072. arXiv:2308.00072 [cs].

J. Keating. Why did One Laptop Per Child fail?, Sept. 2009. URL https://foreignpolicy.com/2009/09/09/why-did-one-laptop-per-child-fail/.

G. Keeling and C. Burr. Digital manipulation and mental integrity. In *The Philosophy of Online Manipulation*, pages 253–271. Routledge, 2022.

G. Keeling and N. Paterson. Proper functions: Etiology without typehood. *Biology & Philosophy*, 37(3):19, 2022.

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of Language Agents, Mar. 2021. URL http://arxiv.org/abs/2103.14659. arXiv:2103.14659 [cs].

Z. Kenton, R. Kumar, S. Farquhar, J. Richens, M. MacDermott, and T. Everitt. Discovering Agents, Aug. 2022. URL http://arxiv.org/abs/2208.08345. arXiv:2208.08345 [cs].

V. Kepuska and G. Bohouta. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103, Las Vegas, NV, Jan. 2018. IEEE. ISBN 9781538646496. doi: 10.1109/CCWC.2018.8301638. URL http://ieeexplore.ieee.org/document/8301638/.

A. Kerasidou. Trust me, I'm a researcher!: The role of trust in biomedical research. *Medicine, Health Care and Philosophy*, 20(1):43–50, Mar. 2017. ISSN 1572-8633. doi: 10.1007/s11019-016-9721-6. URL https://doi.org/10.1007/s11019-016-9721-6.

C. X. Kerasidou, A. Kerasidou, M. Buscher, and S. Wilkinson. Before and beyond trust: reliance in medical AI. *Journal of Medical Ethics*, 48 (11):852–856, Nov. 2022. ISSN 0306-6800, 1473-4257. doi: 10.1136/medethics-2020-107095. URL https://jme.bmj.com/lookup/doi/10.1136/medethics-2020-107095.

S. M. Khan. AI Chips: What They Are and Why They Matter, 2020. URL https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

M. Khwaja, M. Ferrer, J. O. Iglesias, A. A. Faisal, and A. Matic. Aligning daily activities with personality: Towards a recommender system for improving wellbeing. In *Proceedings of the 13th AMC Conference on Recommender Systems*, pages 368–372, 2019.

A. Kierans, H. Hazan, and S. Dori-Hacohen. Quantifying misalignment between agents. *ML Safety@ NeurIPS 2022*, 2022.

Y. Kim and S. S. Sundar. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250, Jan. 2012. ISSN 0747-5632. doi: 10.1016/j.chb.2011.09.006. URL https://www.sciencedirect.com/science/article/pii/S0747563211001993.

E. Kimani, K. Rowan, D. McDuff, M. Czerwinski, and G. Mark. A conversational agent in support of productivity and wellbeing at work. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

W. M. Kincannon, M. Zahn, R. Clare, J. Lusty Beech, A. Romberg, J. Larson, B. Bothner, G. T. Beckham, J. E. McGeehan, and J. L. DuBois. Biochemical and structural characterization of an aromatic ring–hydroxylating dioxygenase for terephthalic acid catabolism. *Proceedings of the National Academy of Sciences*, 119(13):e2121426119, Mar. 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2121426119. URL https://pnas.org/doi/full/10.1073/pnas.2121426119.

M. F. King, V. F. Renó, and E. M. Novo. The concept, dimensions and methods of assessment of human well-being within a socioecological context: A literature review. *Social Indicators Research*, 116:681–698, 2014.

T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, 26(1):89–120, Feb. 2020. ISSN 1471-5546. doi: 10.1007/s11948-018-00081-0. URL https://doi.org/10.1007/s11948-018-00081-0.

P. Kirby. Shadow Schooling: Private tuition and social mobility in the UK. Technical report, The Sutton Trust, Sept. 2016. URL https://www.suttontrust.com/wp-content/uploads/2019/12/Shadow-Schooling-formatted-report_FINAL.pdf.

H. Kirk, A. Bean, B. Vidgen, P. Rottger, and S. Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.148. URL https://aclanthology.org/2023.emnlp-main.148.

O. Kivinen and T. Piiroinen. Epoch-Making Changes in the Cultural Evolution of Communication: Communication technologies seen as organized hubs of skillful human activities. *Journal for the Theory of Social Behaviour*, 53(2):221–237, June 2023. ISSN 0021-8308, 1468-5914. doi: 10.1111/jtsb.12361. URL https://onlinelibrary.wiley.com/doi/10.1111/jtsb.12361.

J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118, June 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2018340118. URL https://pnas.org/doi/full/10.1073/pnas.2018340118.

J. Kleinig and N. G. Evans. Human flourishing, human dignity, and human rights. *Law and Philosophy*, 32(5):539–564, 2013.

M. Klenk. Digital Well-Being and Manipulation Online. In C. Burr and L. Floridi, editors, *Ethics of Digital Well-Being: A Multidisciplinary Approach*, Philosophical Studies Series, pages 81–100. Springer International Publishing, Cham, 2020. ISBN 9783030505851. doi: 10.1007/978-3-030-50585-1_4. URL https://doi.org/10.1007/978-3-030-50585-1_4.

M. Klenk. (Online) manipulation: sometimes hidden, always careless. *Review of Social Economy*, 80(1):85–105, Jan. 2022. ISSN 0034-6764, 1470-1162. doi: 10.1080/00346764.2021.1894350. URL https://www.tandfonline.com/doi/full/10.1080/00346764.2021.1894350.

A. Kling. What Gets Expensive, and Why?, Dec. 2018. URL https://medium.com/@arnoldkling/what-gets-expensive-and-why-33bf4b891be2.

W. Knight. Google Assistant Finally Gets a Generative AI Glow-Up. *Wired*, 2023. ISSN 1059-1028. URL https://www.wired.com/story/google-assistant-multi-modal-upgrade-bard-generative-ai/.

A. B. Kocaballi, J. C. Quiroz, L. Laranjo, D. Rezazadegan, R. Kocielnik, L. Clark, Q. V. Liao, S. Y. Park, R. J. Moore, and A. Miner. Conversational agents for health and wellbeing. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, Apr. 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915768117. URL https://pnas.org/doi/full/10.1073/pnas.1915768117.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large Language Models are Zero-Shot Reasoners. arXiv, 2022. doi: 10.48550/arXiv.2205.11916. URL http://arxiv.org/abs/2205.11916. arXiv:2205.11916 [cs].

M. Komeili, K. Shuster, and J. Weston. Internet-Augmented Dialogue Generation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.579. URL https://aclanthology.org/2022.acl-long.579.pdf.

A. Korinek and J. E. Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence: An agenda*, pages 349–390. University of Chicago Press, 2018.

C. M. Korsgaard. *The Sources of Normativity*. Cambridge University Press, New York, 1996.

R. Koster, J. Balaguer, A. Tacchetti, A. Weinstein, T. Zhu, O. Hauser, D. Williams, L. Campbell-Gillingham, P. Thacker, M. Botvinick, and C. Summerfield. Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*, 6(10):1398–1407, July 2022. ISSN 2397-3374. doi: 10.1038/s41562-022-01383-x. URL https://www.nature.com/articles/s41562-022-01383-x.

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of AI ingenuity, Apr. 2020. URL https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

I. Kramer. The Birth of Privacy Law: A Century Since Warren and Brandeis. *Catholic University Law Review*, 39(3):703–724, Jan. 1990. URL https://scholarship.law.edu/lawreview/vol39/iss3/3.

R. M. Kramer. Rethinking Trust. *Harvard Business Review*, June 2009. ISSN 0017-8012. URL https://hbr.org/2009/06/rethinking-trust.

D. Kreiss and B. Barrett. Democratic tradeoffs: Platforms and political advertising. *Ohio St. Tech. LJ*, 16:493, 2020.

S. Kreps and D. L. Kriner. The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. *New Media & Society*, Mar. 2023. ISSN 1461-4448, 1461-7315. doi: 10.1177/14614448231160526. URL http://journals.sagepub.com/doi/10.1177/14614448231160526.

J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk. Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, Jan. 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00447. URL https://arxiv.org/pdf/2103.12028.pdf. arXiv:2103.12028 [cs].

V. Krstić and C. Saville. Deception (Under Uncertainty) as a Kind of Manipulation. *Australasian Journal of Philosophy*, 97(4):830–835, Oct. 2019. ISSN 0004-8402, 1471-6828. doi: 10.1080/00048402.2019.1604777. URL https://www.tandfonline.com/doi/full/10.1080/00048402.2019.1604777.

A. B. Krueger and D. A. Schkade. The reliability of subjective well-being measures. *Journal of Public Economics*, 92(8-9):1833–1845, 2008.

A. B. Krueger and A. A. Stone. Progress in measuring subjective well-being. *Science*, 346(6205):42–43, 2014.

T.-C. Kuo, C.-Y. Kuo, and L.-W. Chen. Assessing environmental impacts of nanoscale semi-conductor manufacturing from the life cycle assessment perspective. *Resources, Conservation and Recycling*, 182:106289, July 2022. ISSN 09213449. doi: 10.1016/j.resconrec.2022.106289. URL https://linkinghub.elsevier.com/retrieve/pii/S0921344922001379.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial Machine Learning at Scale, Feb. 2017. URL http://arxiv.org/abs/1611.01236. arXiv:1611.01236 [cs, stat].

A. Kurakin, N. Ponomareva, U. Syed, L. MacDermed, and A. Terzis. Harnessing large-language models to generate private synthetic text, June 2023. URL http://arxiv.org/abs/2306.01684. arXiv:2306.01684 [lg, cr].

L. Kvasny. Cultural (Re)production of digital inequality in a US community technology initiative. *Information, Communication & Society*, 9 (2):160–181, Apr. 2006. ISSN 1369-118X, 1468-4462. doi: 10.1080/13691180600630740. URL http://www.tandfonline.com/doi/abs/10.1080/13691180600630740.

O. Kässi, V. Lehdonvirta, and F. Stephany. How many online workers are there in the world? A data-driven assessment. *Open Research Europe*, 1:53, Oct. 2021. ISSN 2732-5121. doi: 10.12688/openreseurope.13639.4. URL https://open-research-europe.ec.europa.eu/articles/1-53/v4.

A. Kääriä. Technology acceptance of voice assistants : anthropomorphism as factor. 2017. URL https://jyx.jyu.fi/handle/123456789/54612.

S. Kühn, T. R. Brick, B. C. N. Müller, and J. Gallinat. Is This Car Looking at You? How Anthropomorphism Predicts Fusiform Face Area Activation when Seeing Cars. *PLoS ONE*, 9(12):e113885, Dec. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0113885. URL https://dx.plos.org/10.1371/journal.pone.0113885.

L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, page 14614448221142007, 2022.

A. Laitinen and O. Sahlgren. Ai systems and respect for human autonomy. *Frontiers in artificial intelligence*, 4:151, 2021.

P. Lally and B. Gardner. Promoting habit formation. *Health Psychology Review*, 7(sup1):S137–S158, 2013.

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

K. C. Land, V. L. Lamb, and S. K. Mustillo. Child and youth well-being in the united states, 1975-1998: Some findings from a new index. *Social Indicators Research*, pages 241–320, 2001.

N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, et al. Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications*, 19:345–359, 2014.

B. Lange, G. Keeling, A. McCroskery, B. Zevenbergen, S. Blascovich, K. Pedersen, A. Lentz, and B. Agüera y Arcas. Engaging engineering teams through moral imagination: a bottom-up approach for responsible innovation and ethical culture change in technology companies. *AI and Ethics*, pages 1–10, 2023.

L. Langosco, J. Koch, L. Sharkey, J. Pfau, L. Orseau, and D. Krueger. Goal Misgeneralization in Deep Reinforcement Learning, Jan. 2023. URL http://arxiv.org/abs/2105.14111. arXiv:2105.14111 [cs].

N. K. Lankton, D. H. McKnight, and J. Tripp. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):1, 2015.

J. A. Lanz. Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity, Apr. 2023. URL https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity.

N. Lanzetti, M. Schiffer, M. Ostrovsky, and M. Pavone. On the Interplay between Self-Driving Cars and Public Transportation, Sept. 2021. URL http://arxiv.org/abs/2109.01627. arXiv:2109.01627 [physics].

P. L. Lanzi and D. Loiacono. ChatGPT and Other Large Language Models as Evolutionary Engines for Online Interactive Collaborative Game Design. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1383–1390, July 2023. doi: 10.1145/3583131.3590351. URL http://arxiv.org/abs/2303.02155. arXiv:2303.02155 [cs].

J. Larson. A Land Full of Gods: Nature Deities in Greek Religion. In D. Ogden, editor, *A Companion to Greek Religion*, pages 56–70. Wiley, 1 edition, Jan. 2007. ISBN 9781405120548 9780470996911. doi: 10.1002/9780470996911.ch4. URL https://onlinelibrary.wiley.com/doi/10.1002/9780470996911.ch4.

S. Larsson and D. R. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4):323–340, Sept. 2000. ISSN 1469-8110, 1351-3249. doi: 10.1017/S1351324900002539. URL https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/information-state-and-dialogue-management-in-the-trindi-dialogue-move-engine-toolkit/3C57FAE606CDC93C172B6090DF0FACE3.

K. Lauter, W. Dai, and K. Laine, editors. *Protecting Privacy through Homomorphic Encryption*. Springer International Publishing, Cham, 2021. ISBN 9783030772864 9783030772871. doi: 10.1007/978-3-030-77287-1. URL https://link.springer.com/10.1007/978-3-030-77287-1.

J. Laux, S. Wachter, and B. Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, page rego.12512, Feb. 2023. ISSN 1748-5983, 1748-5991. doi: 10.1111/rego.12512. URL https://onlinelibrary.wiley.com/doi/10.1111/rego.12512.

H. Law. The Persuasion Game, 2023. URL https://www.learningfromexamples.com/p/the-persuasion-game.

R. Layard. Measuring subjective well-being. *Science*, 327(5965):534–535, 2010.

R. Layard and J.-E. De Neve. *Wellbeing: Science and Policy*. Cembridge University Press, 2023.

S. Lazar. Power and AI: Nature and Justification. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, and B. Zhang, editors, *The Oxford Handbook of AI Governance*. Oxford University Press, 1 edition, May 2022. ISBN 9780197579329 9780197579350. doi: 10.1093/oxfordhb/9780197579329.013.12. URL https://academic.oup.com/edited-volume/41989/chapter/355437737.

S. Lazar. Legitimacy, Authority, and Democratic Duties of Explanation, Oct. 2023. URL http://arxiv.org/abs/2208.08628. arXiv:2208.08628 [cs].

S. Lazar and A. Nelson. AI safety on whose terms? *Science*, 381(6654):138–138, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi8982. URL https://www.science.org/doi/10.1126/science.adi8982.

L. Le Bigot, J.-F. Rouet, and E. Jamet. Effects of Speech- and Text-Based Interaction Modes in Natural Language Human-Computer Dialogue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(6):1045–1053, Dec. 2007. ISSN 0018-7208, 1547-8181. doi: 10.1518/001872007X249901. URL 10.1518/001872007X249901.

K. Lee, O. Firat, A. Agarwal, C. Fannjiang, and D. Sussillo. Hallucinations in Neural Machine Translation, 2019. URL https://openreview.net/forum?id=SkxJ-309FQ.

K. Y. Lee, L. Sheehan, K. Lee, and Y. Chang. The continuation and recommendation intention of artificial intelligence-based voice assistant systems (Aivas): the influence of personal traits. *Internet Research*, 31(5):1899–1939, Nov. 2021. ISSN 1066-2243. doi: 10.1108/INTR-06-2020-0327. URL https://www.emerald.com/insight/content/doi/10.1108/INTR-06-2020-0327/full/html.

M. Lee, M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X. L. Li, F. Ladhak, F. Rong, R. E. Wang, M. Kwon, J. S. Park, H. Cao, T. Lee, R. Bommasani, M. Bernstein, and P. Liang. Evaluating Human-Language Model Interaction, Sept. 2023a. URL http://arxiv.org/abs/2212.09746. arXiv:2212.09746 [cs].

N. Lee and S. Clarke. Do low-skilled workers gain from high-tech employment growth? High-technology multipliers, employment and wages in Britain. *Research Policy*, 48(9):103803, Nov. 2019. ISSN 00487333. doi: 10.1016/j.respol.2019.05.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0048733319301234.

N. Lee, W. Ping, P. Xu, M. Patwary, P. Fung, M. Shoeybi, and B. Catanzaro. Factuality Enhanced Language Models for Open-Ended Text Generation, Mar. 2023b. URL http://arxiv.org/abs/2206.04624. arXiv:2206.04624 [cs].

J. Lehman. Machine Love, Feb. 2023. URL http://arxiv.org/abs/2302.09248. arXiv:2302.09248 [cs] version: 1.

J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. J. Bentley, S. Bernard, G. Beslon, D. M. Bryson, N. Cheney, P. Chrabaszcz, A. Cully, S. Doncieux, F. C. Dyer, K. O. Ellefsen, R. Feldt, S. Fischer, S. Forrest, A. Fŕenoy, C. Gagńe, L. Le Goff, L. M. Grabowski, B. Hodjat, F. Hutter, L. Keller, C. Knibbe, P. Krcah, R. E. Lenski, H. Lipson, R. MacCurdy, C. Maestre, R. Miikkulainen, S. Mitri, D. E. Moriarty, J.-B. Mouret, A. Nguyen, C. Ofria, M. Parizeau, D. Parsons, R. T. Pennock, W. F. Punch, T. S. Ray, M. Schoenauer, E. Schulte, K. Sims, K. O. Stanley, F. Taddei, D. Tarapore, S. Thibault, R. Watson, W. Weimer, and J. Yosinski. The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities. *Artificial Life*, 26(2):274–306, May 2020. ISSN 1064-5462, 1530-9185. doi: 10.1162/artl_a_00319. URL https://direct.mit.edu/artl/article/26/2/274-306/93255.

V. Lei and F. Vesely. In-Group versus Out-Group Trust: The Impact of Income Inequality. *Southern Economic Journal*, 76(4):1049–1063, Apr. 2010. ISSN 0038-4038. doi: 10.4284/sej.2010.76.4.1049. URL http://doi.wiley.com/10.4284/sej.2010.76.4.1049.

J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel. Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research, Mar. 2019. URL http://arxiv.org/abs/1903.00742. arXiv:1903.00742 [cs, q-bio].

J. Leike. Combining weak-to-strong generalization with scalable oversight, 2023. URL https://aligned.substack.com/p/combining-w2sg-with-scalable-oversight.

J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction, Nov. 2018. URL http://arxiv.org/abs/1811.07871. arXiv:1811.07871 [cs, stat].

J. Lenman. Consequentialism and cluelessness. *Philosophy & public affairs*, 29(4):342–370, 2000.

J. Letchford, D. Korzhyk, and V. Conitzer. On the value of commitment. *Autonomous Agents and Multi-Agent Systems*, 28(6):986–1016, Nov. 2014. ISSN 1387-2532, 1573-7454. doi: 10.1007/s10458-013-9246-9. URL http://link.springer.com/10.1007/s10458-013-9246-9.

K. Letheren, J. Jetten, J. Roberts, and J. Donovan. Robots should be seen and not heard… sometimes: Anthropomorphism and ai service robot interactions. *Psychology & Marketing*, 38(12):2393–2406, 2021.

N. G. Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

J. Levin. Human flourishing and population health: Meaning, measurement, and implications. *Perspectives in Biology and Medicine*, 63(3): 401–419, 2020.

B. Levinstein and D. A. Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *arXiv preprint arXiv:2307.00175*, 2023.

B. S. Levy and J. A. Patz. Climate Change, Human Rights, and Social Justice. *Annals of Global Health*, 81(3):310, Nov. 2015. ISSN 2214-9996. doi: 10.1016/j.aogh.2015.08.008. URL https://annalsofglobalhealth.org/articles/10.1016/j.aogh.2015.08.008.

S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, Dec. 2012. ISSN 1529-1006, 1539-6053. doi: 10.1177/1529100612451018. URL http://journals.sagepub.com/doi/10.1177/1529100612451018.

K. Lewicki, M. S. A. Lee, J. Cobbe, and J. Singh. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, Apr. 2023. doi: 10.1145/3544548.3581463. URL http://arxiv.org/abs/2302.01448. arXiv:2302.01448 [cs].

P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk. MLQA: Evaluating Cross-lingual Extractive Question Answering, May 2020. URL http://arxiv.org/abs/1910.07475. arXiv:1910.07475 [cs].

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Apr. 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335, June 2023a. ISSN 25897500. doi: 10.1016/S2589-7500(23)00083-3. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00083-3/fulltext.

J. Li and A. Brar. The use and impact of digital technologies for and on the mental health and wellbeing of indigenous people: a systematic review of empirical studies. *Computers in Human Behavior*, 126:106988, 2022.

P. Li, J. Yang, M. A. Islam, and S. Ren. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models, Oct. 2023b. URL http://arxiv.org/abs/2304.03271. arXiv:2304.03271 [cs].

X. S. Li, S. Kim, K. W. Chan, and A. L. McGill. Detrimental effects of anthropomorphism on the perceived physical safety of artificial agents in dangerous situations. *International Journal of Research in Marketing*, 40(4):841–864, 2023c.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic Evaluation of Language Models, Nov. 2022. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs] version: 1.

Y. Liang. Recommender system for developing new preferences and goals. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 611–615, 2019.

Y. Liang, C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao, Y. Wang, L. Shou, M. Gong, and N. Duan. TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs, Mar. 2023. URL http://arxiv.org/abs/2303.16434. arXiv:2303.16434 [cs].

C. W. Lidz. Coercion in psychiatric care: what have we learned from research? *The Journal of the American Academy of Psychiatry and the Law*, 26(4):631–637, 1998. ISSN 1093-6793.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's Verify Step by Step, May 2023. URL http://arxiv.org/abs/2305.20050. arXiv:2305.20050 [cs].

B. Lim, M. Flageat, and A. Cully. Efficient exploration using model-based quality-diversity with gradients. *arXiv preprint arXiv:2211.12610*, 2022.

L. Lima, V. Furtado, E. Furtado, and V. Almeida. Empirical Analysis of Bias in Voice-based Personal Assistants. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 533–538, San Francisco USA, May 2019. ACM. ISBN 9781450366755. doi: 10.1145/3308560.3317597. URL https://dl.acm.org/doi/10.1145/3308560.3317597.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

J. Lingel and K. Crawford. "Alexa, Tell Me about Your Mother": The History of the Secretary and the End of Secrecy. *Catalyst: Feminism, Theory, Technoscience*, 6(1), May 2020. ISSN 2380-3312. doi: 10.28968/cftt.v6i1.29949. URL https://catalystjournal.org/index.php/catalyst/article/view/29949.

S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan, May 2021. ACM. ISBN 9781450380966. doi: 10.1145/3411764.3445488. URL https://dl.acm.org/doi/10.1145/3411764.3445488.

A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423, May 2020. doi: 10.1109/ICASSP40776.2020.9054458. URL http://arxiv.org/abs/1910.12638. arXiv:1910.12638 [cs, eess].

S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023a.

X. Liu, W. Ai, H. Li, J. Tang, G. Huang, F. Feng, and Q. Mei. Deriving user preferences of mobile apps from their management activities. *ACM Transactions on Information Systems (TOIS)*, 35(4):1–32, 2017.

Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study, May 2023b. URL https://arxiv.org/pdf/2305.13860.pdf. arXiv:2305.13860 [cs].

N. Lomas. Social media a factor in death of uk schoolgirl, inquest finds. https://techcrunch.com/2022/09/30/molly-russell-inquest-verdict/, 2022. Accessed: 2023-07-13.

J. Long. Large language model guided tree-of-thought, 2023.

H. Longino. Feminist standpoint theory and the problems of knowledge. *Signs: Journal of Women in Culture and Society*, 19(1):201–212, 1993. doi: 10.1086/494867.

S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno, and D. Ippolito. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity, May 2023. URL https://arxiv.org/pdf/2305.13169.pdf. arXiv:2305.13169 [cs].

V. Lotz, A. C. Valdez, and M. Ziefle. Don't Stand so Close to Me: Acceptance of Delegating Intimate Health Care Tasks to Assistive Robots. In V. G. Duffy, M. Ziefle, P.-L. P. Rau, and M. M. Tseng, editors, *Human-Automation Interaction*, volume 12, pages 3–21. Springer International Publishing, Cham, 2023. ISBN 9783031107870 9783031107887. doi: 10.1007/978-3-031-10788-7_1. URL https://link.springer.com/10.1007/978-3-031-10788-7_1.

S. Lovato and A. M. Piper. "Siri, is this you?": Understanding young children's interactions with voice input systems. In *Proceedings of the 14th International Conference on Interaction Design and Children*, pages 335–338, Boston Massachusetts, June 2015. ACM. ISBN 9781450335904. doi: 10.1145/2771839.2771910. URL https://dl.acm.org/doi/10.1145/2771839.2771910.

P.-F. Lovens. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là". *La Libre*, Mar. 2023. URL https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/.

A. S. Luccioni and A. Hernandez-Garcia. Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning, Feb. 2023. URL http://arxiv.org/abs/2302.08476. arXiv:2302.08476 [cs].

A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, Nov. 2022. URL http://arxiv.org/abs/2211.02001. arXiv:2211.02001 [cs].

R. D. Luce and H. Raiffa. *Games and decisions: Introduction and critical survey*. Games and decisions: Introduction and critical survey. Wiley, Oxford, England, 1957.

E. Ludlow, P. Anand, and S. McBride. Character.AI in Early Talks for Funding at More Than $5 Billion Valuation. *Bloomberg*, Sept. 2023. URL https://www.bloomberg.com/news/articles/2023-09-28/character-ai-in-early-talks-for-funding-at-more-than-5-billion-valuation.

N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin. Analyzing Leakage of Personally Identifiable Information in Language Models, Apr. 2023. URL http://arxiv.org/abs/2302.00539. arXiv:2302.00539 [cs].

K. Lum and W. Isaac. To Predict and Serve? *Significance*, 13(5):14–19, Oct. 2016. ISSN 1740-9705, 1740-9713. doi: 10.1111/j.1740-9713.2016.00960.x. URL https://academic.oup.com/jrssig/article/13/5/14/7029190.

J. Luo, C. Paduraru, O. Voicu, Y. Chervonyi, S. Munns, J. Li, C. Qian, P. Dutta, J. Q. Davis, N. Wu, et al. Controlling commercial cooling systems using reinforcement learning. *arXiv preprint arXiv:2211.07357*, 2022.

D. Lyon. Biometrics, Identification, and Surveillance. *Bioethics*, 22(9):499–508, Nov. 2008. ISSN 0269-9702, 1467-8519. doi: 10.1111/j.1467-8519.2008.00697.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8519.2008.00697.x.

M. M. Maas. Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment, Dec. 2018. URL https://papers.ssrn.com/abstract=3756941.

A. Macdonald. India okays 22 financial entities to perform Aadhaar biometric authentication | Biometric Update, May 2023. URL https://www.biometricupdate.com/202305/india-okays-22-financial-entities-to-perform-aadhaar-biometric-authentication.

S. Machkovech. Report: Facebook helped advertisers target teens who feel "worthless" [Updated], May 2017. URL https://arstechnica.com/information-technology/2017/05/facebook-helped-advertisers-target-teens-who-feel-worthless/.

A. MacKenzie and I. Bhatt. Lies, Bullshit and Fake News: Some Epistemological Concerns. *Postdigital Science and Education*, 2(1):9–13, Jan. 2020. ISSN 2524-4868. doi: 10.1007/s42438-018-0025-4. URL https://doi.org/10.1007/s42438-018-0025-4.

C. Mackenzie and N. Stoljar. *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press, 2000.

C. Mackenzie, W. Rogers, and S. Dodds, editors. *Vulnerability: New Essays in Ethics and Feminist Philosophy*. Oxford University Press, Dec. 2013. ISBN 9780199316649. doi: 10.1093/acprof:oso/9780199316649.001.0001. URL https://academic.oup.com/book/1543.

R. Macklin. Man, Mind, and Morality: The Ethics of Behavior Control, 1982. URL https://repository.library.georgetown.edu/handle/10822/792035.

P. Madhavan and D. A. Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

J. Makhoul, F. Jelinek, L. Rabiner, C. Weinstein, and V. Zue. White Paper on Spoken Language Systems. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989. URL https://aclanthology.org/H89-2077.

J. Manyika and K. Sneader. AI, automation, and the future of work: Ten things to solve for. Technical report, McKinsey Global Institute, June 2018. URL https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for.

J. Manyika and M. Spence. The Coming AI Economic Revolution: Can Artificial Intelligence Reverse the Productivity Slowdown? *Foreign Aff.*, 102:70, 2023.

V. Marda and S. Narayan. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence*, 3(3):187–189, Mar. 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00323-0. URL https://www.nature.com/articles/s42256-021-00323-0.

N. Marks. *The unhappy planet index: An index of human well-being and environmental impact*. New Economics Foundation, 2006.

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

B. Marr. Microsoft's Plan To Infuse AI And ChatGPT Into Everything, 2023. URL https://www.forbes.com/sites/bernardmarr/2023/03/06/microsofts-plan-to-infuse-ai-and-chatgpt-into-everything/.

K. Marshall, A. Thieme, J. Wallace, J. Vines, G. Wood, and M. Balaam. Making wellbeing: A process of user-centered design. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, pages 755–764, 2014.

J. L. Martin and K. E. Wright. Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, 44(4): 613–630, Aug. 2023. ISSN 0142-6001, 1477-450X. doi: 10.1093/applin/amac066. URL https://academic.oup.com/applij/article/44/4/613/6901317.

S. Martin and J. Marks. *Messengers: Who We Listen To, Who We Don't, And Why*. Random House, Sept. 2019. ISBN 9781473560727. Google-Books-ID: RVJfDwAAQBAJ.

D. Martin Jr, V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572*, 2020.

J. Martínez Torres, C. Iglesias Comesaña, and P. J. García-Nieto. Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, 10(10):2823–2836, Oct. 2019. ISSN 1868-8071, 1868-808X. doi: 10.1007/s13042-018-00906-1. URL http://link.springer.com/10.1007/s13042-018-00906-1.

G. Marwell and P. Oliver. *The Critical Mass in Collective Action*. Cambridge University Press, 1 edition, Mar. 1993. ISBN 9780521308397 9780521039550 9780511663765. doi: 10.1017/CBO9780511663765. URL https://www.cambridge.org/core/product/identifier/9780511663765/type/book.

C. N. Mascie-Taylor and E. Karim. The burden of chronic disease. *Science*, 302(5652):1921–1922, 2003.

D. S. Massey and N. A. Denton. *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press, 1993. ISBN 9780674018211. Google-Books-ID: uGslMsIBNBsC.

N. Mattis, P. Masur, J. Möller, and W. van Atteveldt. Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *New Media & Society*, page 146144482211044, June 2022. ISSN 1461-4448, 1461-7315. doi: 10.1177/14614448221104413. URL http://journals.sagepub.com/doi/10.1177/14614448221104413.

R. C. Mayer, J. H. Davis, and F. D. Schoorman. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3): 709, July 1995. ISSN 03637425. doi: 10.2307/258792. URL http://www.jstor.org/stable/258792?origin=crossref.

R. Mayrhofer, J. V. Stoep, C. Brubaker, and N. Kralevich. The Android Platform Security Model, Dec. 2020. URL http://arxiv.org/abs/1904.05572. arXiv:1904.05572 [cs].

D. J. McAllister. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, 38(1):24–59, Feb. 1995. ISSN 0001-4273, 1948-0989. doi: 10.2307/256727. URL http://amj.aom.org/cgi/doi/10.2307/256727.

V. McArthur. Communication technologies and cultural identity a critical discussion of icts for development. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, pages 910–914. IEEE, 2009.

H. J. McCloskey. Privacy and the Right to Privacy. *Philosophy*, 55(211):17–38, Jan. 1980. ISSN 0031-8191, 1469-817X. doi: 10.1017/S0031819100063725. URL https://www.cambridge.org/core/product/identifier/S0031819100063725/type/journal_article.

M. McGillivray. Human well-being: Issues, concepts and measures. In *Human well-being: Concept and measurement*, pages 1–22. Springer, 2007.

E. McGinnies and C. D. Ward. Better Liked than Right: Trustworthiness and Expertise as Factors in Credibility. *Personality and Social Psychology Bulletin*, 6(3):467–472, Sept. 1980. ISSN 0146-1672, 1552-7433. doi: 10.1177/014616728063023. URL http://journals.sagepub.com/doi/10.1177/014616728063023.

J. A. McGregor. Reconciling universal frameworks and local realities in understanding and measuring wellbeing. In I. Bache and K. Scott, editors, *The Politics of Wellbeing: Theory, policy and practice*, pages 197–224. Springer, 2018.

S. McGregor. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):15458–15463, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i17.17817. URL https://ojs.aaai.org/index.php/AAAI/article/view/17817.

K. R. McKee, X. Bai, and S. Fiske. Humans perceive warmth and competence in artificial intelligence. preprint, PsyArXiv, Feb. 2021. URL https://osf.io/5ursp.

K. R. McKee, E. Hughes, T. O. Zhu, M. J. Chadwick, R. Koster, A. G. Castaneda, C. Beattie, T. Graepel, M. Botvinick, and J. Z. Leibo. A multi-agent reinforcement learning model of reputation and cooperation in human groups, Feb. 2023a. URL http://arxiv.org/abs/2103.04982. arXiv:2103.04982 [cs].

K. R. McKee, A. Tacchetti, M. A. Bakker, J. Balaguer, L. Campbell-Gillingham, R. Everett, and M. Botvinick. Scaffolding cooperation in human groups with deep reinforcement learning. *Nature Human Behaviour*, 7(10):1787–1796, Sept. 2023b. ISSN 2397-3374. doi: 10.1038/s41562-023-01686-7. URL https://www.nature.com/articles/s41562-023-01686-7.

R. McQuaid. Youth unemployment produces multiple scarring effects, Feb. 2017. URL https://blogs.lse.ac.uk/europpblog/2017/02/18/youth-unemployment-scarring-effects/.

M. McTear. *Conversational AI: dialogue systems, conversational agents, and chatbots*. Number #48 in Synthesis lectures on human language technologies. Morgan & Claypool Publishers, San Rafael, California, 2021. ISBN 9781636390338 9781636390314 9781636390321. URL https://www.amazon.co.uk/Conversational-AI-Dialogue-Systems-Chatbots/dp/1636390331/ref=tmm_hrd_swatch_0?_encoding=UTF8&qid=1688737381&sr=1-1.

J. Meng and Y. N. Dai. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication*, 26(4):207–222, Sept. 2021. ISSN 1083-6101. doi: 10.1093/jcmc/zmab005. URL https://academic.oup.com/jcmc/article/26/4/207/6278042.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT, Jan. 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*, 4:725911, Nov. 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.725911. URL https://www.frontiersin.org/articles/10.3389/frai.2021.725911/full.

K. Merrill, J. Kim, and C. Collins. AI companions for lonely individuals and the role of social presence. *Communication Research Reports*, 39 (2):93–103, Mar. 2022. ISSN 0882-4096, 1746-4099. doi: 10.1080/08824096.2022.2045929. URL https://www.tandfonline.com/doi/full/10.1080/08824096.2022.2045929.

G. S. Mesch and I. Talmud. Ethnic Differences in Internet Access: The role of occupation and exposure. *Information, Communication & Society*, 14(4):445–471, June 2011. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2011.562218. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2011.562218.

Meta. Introducing New AI Experiences Across Our Family of Apps and Devices, Sept. 2023. URL https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/.

Meta Fundamental AI Research Diplomacy Team, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, M. Komeili, K. Konath, M. Kwon, A. Lerer, M. Lewis, A. H. Miller, S. Mitts, A. Renduchintala, S. Roller, D. Rowe, W. Shi, J. Spisak, A. Wei, D. Wu, H. Zhang, and M. Zijlstra. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, Dec. 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/10.1126/science.ade9097.

T. Metzinger. EU guidelines: Ethics washing made in Europe. *Der Tagesspiegel*, Apr. 2019. ISSN 1865-2263. URL https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html.

G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom. Augmented Language Models: a Survey, Feb. 2023. URL http://arxiv.org/abs/2302.07842. arXiv:2302.07842 [cs].

F. Michaud, T. Salter, A. Duquette, H. Mercier, M. Lauria, H. Larouche, and F. Larose. Assistive Technologies and Children-Robot Interaction. pages 45–49, Jan. 2007.

H. Mieczkowski, J. T. Hancock, M. Naaman, M. Jung, and J. Hohenstein. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–14, Apr. 2021. ISSN 2573-0142. doi: 10.1145/3449091. URL https://dl.acm.org/doi/10.1145/3449091.

A. Milanez. The Impact of AI on the Workplace: Evidence from OECD Case Studies of AI Implementation. OECD Social, Employment and Migration Working Papers 289, OECD, Mar. 2023. URL https://www.oecd-ilibrary.org/social-issues-migration-health/the-impact-of-ai-on-the-workplace-evidence-from-oecd-case-studies-of-ai-implementation_2247ce58-en.

S. Milano, M. Taddeo, and L. Floridi. Recommender systems and their ethical challenges. *AI & SOCIETY*, 35(4):957–967, Dec. 2020. ISSN 1435-5655. doi: 10.1007/s00146-020-00950-y. URL https://doi.org/10.1007/s00146-020-00950-y.

S. Milano, B. Mittelstadt, S. Wachter, and C. Russell. Epistemic fragmentation poses a threat to the governance of online targeting. *Nature Machine Intelligence*, 3(6):466–472, June 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00358-3. URL https://www.nature.com/articles/s42256-021-00358-3.

P. R. Milgrom and J. Roberts. *Economics, Organization, and Management*. Prentice-Hall, 1992. ISBN 9780132246507. Google-Books-ID: cCi3AAAAIAAJ.

J. S. Mill. *On Liberty and Other Essays*. OUP Oxford, Mar. 1998. ISBN 9780191611087. Google-Books-ID: vcSFmFnICk0C original-date: 1859.

C. J. Mills. *Influence: Coercion, manipulation, and persuasion*. PhD thesis, Princeton University, 1991. URL https://www.proquest.com/openview/6193d9f104072d14d2a8f10818604aa1/1?pq-origsite=gscholar&cbl=18750&diss=y.

D. Milmo and D. Anguiano. Facebook, Instagram and WhatsApp working again after global outage took down platforms. *The Guardian*, Oct. 2021. ISSN 0261-3077. URL https://www.theguardian.com/technology/2021/oct/04/facebook-instagram-and-whatsapp-hit-by-outage.

M. Milossi, E. Alexandropoulou-Egyptiadou, and K. E. Psannis. AI Ethics: Algorithmic Determinism or Self-Determination? The GPDR Approach. *IEEE Access*, 9:58455–58466, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3072782. URL https://ieeexplore.ieee.org/document/9400809/.

B. H. Min and C. Borch. Systemic failures and organizational risk management in algorithmic trading: Normal accidents and high reliability in financial markets. *Social Studies of Science*, 52(2):277–302, Apr. 2022. ISSN 0306-3127, 1460-3659. doi: 10.1177/03063127211048515. URL http://journals.sagepub.com/doi/10.1177/03063127211048515.

M. Mingus. Changing the Framework: Disability Justice, Feb. 2011. URL https://leavingevidence.wordpress.com/2011/02/12/changing-the-framework-disability-justice/.

M. Mingus. Access Intimacy, Interdependence and Disability Justice, Apr. 2017. URL https://leavingevidence.wordpress.com/2017/04/12/access-intimacy-interdependence-and-disability-justice/.

M. Miringoff and M.-L. Miringoff. *The social health of the nation: How America is really doing*. Oxford University Press, 1999.

R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht. A Survey of Ad Hoc Teamwork Research, Aug. 2022. URL http://arxiv.org/abs/2202.10450. arXiv:2202.10450 [cs].

M. U. Mirza, A. Richter, E. H. Van Nes, and M. Scheffer. Technology driven inequality leads to poverty and resource depletion. *Ecological Economics*, 160:215–226, June 2019. ISSN 09218009. doi: 10.1016/j.ecolecon.2019.02.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0921800918306542.

P. Mishkin, L. Ahmad, M. Brundage, G. Krueger, and G. Sastry. DALL·E 2 Preview – Risks and Limitations, 2022. URL https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

M. Mitchell. How do we know how smart AI systems are? *Science*, 381(6654):adj5957, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adj5957. URL https://www.science.org/doi/10.1126/science.adj5957.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, Atlanta GA USA, Jan. 2019. ACM. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://dl.acm.org/doi/10.1145/3287560.3287596.

C. Mitelut, B. Smith, and P. Vamplew. Intent-aligned AI systems deplete human agency: The need for agency foundations research in AI safety, 2023.

S. Mithen and P. Boyer. Anthropomorphism and the evolution of cognition. *Journal of the Royal Anthropological Institute*, 2(4):717–722, Dec. 1996. ISSN 13590987. URL https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=13590987&v=2.1&it=r&id=GALE%7CA19225795&sid=googleScholar&linkaccess=abs.

MITRE. MITRE | ATLAS™. URL https://atlas.mitre.org/.

B. Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507, Nov. 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0114-4. URL https://www.nature.com/articles/s42256-019-0114-4.

B. Mittelstadt, S. Wachter, and C. Russell. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default, Mar. 2023. URL https://arxiv.org/pdf/2302.02404.pdf. arXiv:2302.02404 [cs].

A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, Oct. 2022. ISSN 1932-4553, 1941-0484. doi: 10.1109/JSTSP.2022.3207050. URL https://arxiv.org/pdf/2205.10643.pdf. arXiv:2205.10643 [cs, eess].

S. Mohamed, M.-T. Png, and W. Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, Dec. 2020. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-020-00405-8. URL http://arxiv.org/abs/2007.04068. arXiv:2007.04068 [cs, stat].

A. Mok. The cofounder of Google's AI division DeepMind says everybody will have their own AI-powered 'chief of staff' over the next five years, Sept. 2023. URL https://www.businessinsider.com/google-deepmind-cofounder-mustafa-suleyman-everyone-will-have-ai-assistant-2023-9.

J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, pages 1–31, 2023.

A. Molnar, G. Miron, M. Barbour, L. Huerta, S. Shafer, J. Rice, A. Glover, N. Browning, S. Hagle, and F. Boninger. *Virtual Schools 2021*. 2021.

A. Monge Roffarello and L. De Russis. The race towards digital wellbeing: Issues and opportunities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

J. G. Monroe and R. A. Williamson. *They dance in the sky: Native American star myths*. Houghton Mifflin, Boston, 1987. ISBN 9780395399705.

M. L. Montagnani and M. Verstraete. What makes data personal? *UC Davis L. Rev.*, 56:1165, 2022.

J. H. Moor. What is Computer Ethics?*. *Metaphilosophy*, 16(4):266–275, Oct. 1985. ISSN 0026-1068, 1467-9973. doi: 10.1111/j.1467-9973.1985.tb00173.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9973.1985.tb00173.x.

M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, Apr. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4. URL https://www.nature.com/articles/s41586-023-05881-4.

T. Moran. Atlantic Plaza Towers tenants won a halt to facial recognition in their building: Now they're calling on a moratorium on all residential use, Jan. 2020. URL https://ainowinstitute.org/publication/atlantic-plaza-towers-tenants-won-a-halt-to-facial-recognition-in-their-building-now-theyre.

E. Moretti. *The new geography of jobs*. Mariner Books/Houghton Mifflin Harcourt, Boston, Mass, 1st mariner books ed edition, 2013. ISBN 9780544028050. OCLC: ocn826453786.

Morgan Stanley. The Surprising Case for Stronger E-commerce Growth, 2023. URL https://www.morganstanley.com/ideas/global-ecommerce-growth-forecast-2022.

M. Mori, K. MacDorman, and N. Kageki. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, June 2012. ISSN 1070-9932. doi: 10.1109/MRA.2012.2192811. URL http://ieeexplore.ieee.org/document/6213238/.

M. R. Morris. AI and accessibility. *Communications of the ACM*, 63(6):35–37, May 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3356727. URL https://dl.acm.org/doi/10.1145/3356727.

D. Mota-Rojas, C. Mariti, A. Zdeinert, G. Riggio, P. Mora-Medina, A. del Mar Reyes, A. Gazzano, A. Domínguez-Oliva, K. Lezama-García, N. José-Pérez, et al. Anthropomorphism and its adverse effects on the distress and welfare of companion animals. *Animals*, 11(11): 3263, 2021.

N. Y. Motlagh, M. Khajavi, A. Sharifi, and M. Ahmadi. The Impact of Artificial Intelligence on the Evolution of Digital Education: A Comparative Study of OpenAI Text Generation Tools including ChatGPT, Bing Chat, Bard, and Ernie, Sept. 2023. URL http://arxiv.org/abs/2309.02029. arXiv:2309.02029 [cs].

J. A. Mourey, J. G. Olson, and C. Yoon. Products as Pals: Engaging with Anthropomorphic Products Mitigates the Effects of Social Exclusion. *Journal of Consumer Research*, page ucx038, Jan. 2017. ISSN 0093-5301, 1537-5277. doi: 10.1093/jcr/ucx038. URL https://academic.oup.com/jcr/article-lookup/doi/10.1093/jcr/ucx038.

M. Mousavi, H. Davulcu, M. Ahmadi, R. Axelrod, R. Davis, and S. Atran. Effective Messaging on Social Media: What Makes Online Content Go Viral? In *Proceedings of the ACM Web Conference 2022*, pages 2957–2966, Virtual Event, Lyon France, Apr. 2022. ACM. ISBN 9781450390965. doi: 10.1145/3485447.3512016. URL https://dl.acm.org/doi/10.1145/3485447.3512016.

S. Moussawi. User Experiences with Personal Intelligent Agents: A Sensory, Physical, Functional and Cognitive Affordances View. In *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research*, pages 86–92, Buffalo-Niagara Falls NY USA, June 2018. ACM. ISBN 9781450357685. doi: 10.1145/3209626.3209709. URL https://dl.acm.org/doi/10.1145/3209626.3209709.

S. Moussawi and R. Benbunan-Fich. The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology*, 40(15):1603–1626, 2021.

S. Moussawi, M. Koufaris, and R. Benbunan-Fich. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets*, 31:343–364, 2021.

P. Mozur. A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*, Oct. 2018. ISSN 0362-4331. URL https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

C. Muller and J. P. d. V. Aguiar. What Is the Digital Divide?, Mar. 2022. URL https://www.internetsociety.org/blog/2022/03/what-is-the-digital-divide/.

T. Munyer and X. Zhong. DeepTextMark: Deep Learning based Text Watermarking for Detection of Large Language Model Generated Text, May 2023. URL http://arxiv.org/abs/2305.05773. arXiv:2305.05773 [cs].

H. Murphy and C. Criddle. Meta prepares chatbots with personas to try to retain users. *Financial Times*, Aug. 2023.

A. Myers. AI's Powers of Political Persuasion, Feb. 2023. URL https://hai.stanford.edu/news/ais-powers-political-persuasion.

S. Myers West. General purpose AI poses serious risks, should not be excluded from the EU's AI Act | policy brief. https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act, 2023.

D. Nahavandi, R. Alizadehsani, A. Khosravi, and U. R. Acharya. Application of artificial intelligence in wearable devices: Opportunities and challenges. *Computer Methods and Programs in Biomedicine*, 213:106541, 2022.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability, Oct. 2023. URL http://arxiv.org/abs/2301.05217. arXiv:2301.05217 [cs].

A. Narayanan. How to recognize AI snake oil, Jan. 2021. URL https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

A. Narayanan and S. Kapoor. GPT-4 and professional benchmarks: the wrong answer to the wrong question, Mar. 2023. URL https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

L. Narens and B. Skyrms. *The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility*. Oxford University Press, 2020.

C. Nass, J. Steuer, E. Tauber, and H. Reeder. Anthropomorphism, agency, and ethopoeia: computers as social actors. In *INTERACT '93 and CHI '93 conference companion on Human factors in computing systems - CHI '93*, pages 111–112, Amsterdam, The Netherlands, 1993. ACM Press. ISBN 9780897915748. doi: 10.1145/259964.260137. URL http://portal.acm.org/citation.cfm?doid=259964.260137.

C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *Conference companion on Human factors in computing systems - CHI '94*, page 204, Boston, Massachusetts, United States, 1994. ACM Press. ISBN 9780897916516. doi: 10.1145/259963.260288. URL http://portal.acm.org/citation.cfm?doid=259963.260288.

National Education Union. State of education: workload and wellbeing. Technical report, National Education Union, Apr. 2023. URL https://neu.org.uk/latest/press-releases/state-education-workload-and-wellbeing.

J. Newman. A Taxonomy of AI Trustworthiness, 2023. URL https://cltc.berkeley.edu/wp-content/uploads/2023/01/TaxonomyofAITrustworthiness.pdf.

K. H. Ngamaba, M. Panagioti, and C. J. Armitage. Income inequality and subjective well-being: a systematic review and meta-analysis. *Quality of Life Research*, 27(3):577–596, Mar. 2018. ISSN 0962-9343, 1573-2649. doi: 10.1007/s11136-017-1719-x. URL http://link.springer.com/10.1007/s11136-017-1719-x.

C. T. Nguyen. Transparency is surveillance. *Philosophy and Phenomenological Research*, 105(2):331–361, 2022. doi: https://doi.org/10.1111/phpr.12823. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/phpr.12823.

T. A. Nguyen, M. Do, A. E. Gerevini, I. Serina, B. Srivastava, and S. Kambhampati. Generating diverse plans to handle unknown and partially known user preferences. *Artificial Intelligence*, 190:1–31, 2012.

P. J. Nickel, M. Franssen, and P. Kroes. Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy*, 23 (3-4):429–444, 2010. doi: 10.1007/s12130-010-9124-6.

L. Nicoletti and D. Bass. Humans Are Biased. Generative AI Is Even Worse. *Bloomberg*, 2023. URL https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

C. P. Niemiec. Eudaimonic well-being. *Encyclopedia of quality of life and well-being research*, pages 2004–2005, 2014.

S. J. Nightingale and H. Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, Feb. 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2120481119. URL https://pnas.org/doi/full/10.1073/pnas.2120481119.

H. Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79(1):119, Feb. 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10.

S. U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, Dec. 2020. ISBN 9781479833641. doi: 10.18574/nyu/9781479833641.001.0001. URL https://www.degruyter.com/document/doi/10.18574/nyu/9781479833641.001.0001/html.

N. Noddings. *Caring: A Relational Approach to Ethics and Moral Education*. University of California Press, 2 edition, 2013. ISBN 9780520275706. URL https://www.jstor.org/stable/10.1525/j.ctt7zw1nb.

R. Noggle. Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly*, 33(1):43–55, 1996. ISSN 0003-0481. URL https://www.jstor.org/stable/20009846.

R. Noggle. Manipulation, salience, and nudges. *Bioethics*, 32(3):164–170, Mar. 2018. ISSN 1467-8519. doi: 10.1111/bioe.12421.

R. Noggle. The Ethics of Manipulation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022. URL https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/.

B. Nolan. The latest version of ChatGPT told a TaskRabbit worker it was visually impaired to get help solving a CAPTCHA, OpenAI test shows, 2023. URL https://www.businessinsider.com/gpt4-openai-chatgpt-taskrabbit-tricked-solve-captcha-test-2023-3.

G. Noone. 'Foundation models' may be the future of AI. They're also deeply flawed, Nov. 2021. URL https://techmonitor.ai/technology/ai-and-automation/foundation-models-may-be-future-of-ai-theyre-also-deeply-flawed.

R. M. Nouh, H.-H. Lee, W.-J. Lee, and J.-D. Lee. A smart recommender based on hybrid learning methods for personal well-being services. *Sensors*, 19(2):431, 2019.

J. Novikova, O. Dušek, and V. Rieser. RankME: Reliable Human Ratings for Natural Language Generation. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2012. URL https://aclanthology.org/N18-2012.

NowPow. NowPow Sign In. URL https://uniteus.com/nowpow-login/.

S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh2586. URL https://www.science.org/doi/10.1126/science.adh2586.

R. Nozick. *Anarchy, state, and utopia*. John Wiley & Sons, 1974.

J. A. Obar and A. Oeldorf-Hirsch. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, Jan. 2020. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2018.1486870. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1486870.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, Oct. 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/10.1126/science.aax2342.

M. O'Brien. Is Bing too belligerent? Microsoft looks to tame AI chatbot. *AP News*, Feb. 2023. URL https://apnews.com/article/technology-science-microsoft-corp-business-software-fb49e5d625bf37be0527e5173116bef3.

OECD. Compendium of OECD well-being indicators, 2011.

OECD. How Good is Your Job? Measuring and Assessing Job Quality. Technical report, OECD, Feb. 2016. URL https://www.oecd.org/sdd/labour-stats/Job-quality-OECD.pdf.

OECD. Bridging the Digital Gender Divide: Include, Upskill, Innovate. Technical report, OECD, 2018a. URL https://www.oecd.org/digital/bridging-the-digital-gender-divide.pdf.

OECD. *Equity in Education: Breaking Down Barriers to Social Mobility*. PISA. OECD, Oct. 2018b. ISBN 9789264056732 9789264073234. doi: 10.1787/9789264073234-en. URL https://www.oecd-ilibrary.org/education/equity-in-education_9789264073234-en.

OECD. *Skills Matter: Additional Results from the Survey of Adult Skills*. OECD Skills Studies. OECD, Nov. 2019. ISBN 9789264604667 9789264811072 9789264799004 9789264332829. doi: 10.1787/1f029d8f-en. URL https://www.oecd-ilibrary.org/education/skills-matter_1f029d8f-en.

OECD. Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems. OECD Digital Economy Papers 312, June 2021. URL https://www.oecd-ilibrary.org/science-and-technology/tools-for-trustworthy-ai_008232ec-en.

OECD. *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*. Educational Research and Innovation. OECD, Mar. 2023. ISBN 9789264451377 9789264765153 9789264639676 9789264920378. doi: 10.1787/73105f99-en. URL https://www.oecd-ilibrary.org/education/is-education-losing-the-race-with-technology_73105f99-en.

C. Oesterheld and V. Conitzer. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36(2): 46, Aug. 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09574-6. URL https://doi.org/10.1007/s10458-022-09574-6.

K. Ognyanova, D. Lazer, R. E. Robertson, and C. Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, June 2020. doi: 10.37016/mr-2020-024. URL https://misinforeview.hks.harvard.edu/?p=1689.

C. L. Oguego, J. C. Augusto, A. Muñoz, and M. Springett. Using argumentation to manage users' preferences. *Future Generation Computer Systems*, 81:235–243, 2018.

S. Oishi, S. Kesebir, and E. Diener. Income Inequality and Happiness. *Psychological Science*, 22(9):1095–1100, Sept. 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797611417262. URL http://journals.sagepub.com/doi/10.1177/0956797611417262.

J. J. Ojeda-Castelo, M. D. L. M. Capobianco-Uriarte, J. A. Piedra-Fernandez, and R. Ayala. A Survey on Intelligent Gesture Recognition Techniques. *IEEE Access*, 10:87135–87156, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3199358. URL https://ieeexplore.ieee.org/document/9858153/.

P. O'Kane, S. Sezer, and K. McLaughlin. Obfuscation: The Hidden Malware. *IEEE Security & Privacy*, 9(5):41–47, Sept. 2011. ISSN 1558-4046. doi: 10.1109/MSP.2011.98. URL https://ieeexplore.ieee.org/abstract/document/5975134.

S. Okasha. *Agents and goals in evolution*. Oxford University Press, 2018.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3): 10.23915/distill.00024.001, Mar. 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

M. Olson Jr. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard Economic Studies. Harvard University Press, Cambridge, MA, Jan. 1965. ISBN 9780674537514.

C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. In-context Learning and Induction Heads, Mar. 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. publisher: Anthropic.

S. M. Omohundro. The Basic AI Drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 2008. URL https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.393.8356.

O. O'Neill. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge University Press, New York, 1989.

O. O'Neill. *Autonomy and Trust in Bioethics*. Cambridge University Press, 1 edition, Apr. 2002. ISBN 9780521815406 9780521894531 9780511606250. doi: 10.1017/CBO9780511606250. URL https://www.cambridge.org/core/product/identifier/9780511606250/type/book.

H. Ono and M. Zavodny. Immigrants, English Ability and the Digital Divide. *Social Forces*, 86(4):1455–1479, June 2008. ISSN 0037-7732, 1534-7605. doi: 10.1353/sof.0.0052. URL https://academic.oup.com/sf/article-lookup/doi/10.1353/sof.0.0052.

OpenAI. Gpt-4v(ision) system card. 2023a. URL https://api.semanticscholar.org/CorpusID:263218031.

OpenAI. ChatGPT plugins, Mar. 2023b. URL https://openai.com/blog/chatgpt-plugins.

OpenAI. Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk, Jan. 2023c. URL https://openai.com/research/forecasting-misuse.

OpenAI. GPT-4 System Card, Mar. 2023d. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf. publisher: OpenAI.

OpenAI. GPT-4 Technical Report, Mar. 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

OpenAI. Technical Requirements for Using ChatGPT, Nov. 2023. URL https://chatgptdetector.co/chat-gpt-requirements/.

A. Orben and A. K. Przybylski. The association between adolescent well-being and digital technology use. *Nature human behaviour*, 3(2): 173–182, 2019.

C. Orpen. Attitude Similarity, Attraction, and Decision-Making in the Employment Interview. *The Journal of Psychology*, 117(1):111–120, May 1984. ISSN 0022-3980, 1940-1019. doi: 10.1080/00223980.1984.9923666. URL http://www.tandfonline.com/doi/abs/10.1080/00223980.1984.9923666.

E. Ostrom. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives*, 14(3):137–158, Sept. 2000. ISSN 0895-3309. doi: 10.1257/jep.14.3.137. URL https://www.aeaweb.org/articles?id=10.1257/jep.14.3.137.

E. Ostrom. A Multi-Scale Approach to Coping with Climate Change and Other Collective Action Problems. *Solutions*, 1:27–36, 2010. URL https://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/5774/A%20Multi-Scale%20Approach%20to%20C...pdf?sequence=1.

M. Otsuka. Prioritarianism and the measure of utility. *Journal of Political Philosophy*, 23(1):1–22, 2015.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

A. Ovadya and J. Whittlestone. Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning, July 2019. URL http://arxiv.org/abs/1907.11274. arXiv:1907.11274 [cs].

A. Ovalle, A. Subramonian, V. Gautam, G. Gee, and K.-W. Chang. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness, July 2023. URL http://arxiv.org/abs/2303.17555. arXiv:2303.17555 [cs].

R. Owen, M. Pansera, P. Macnaghten, and S. Randles. Organisational institutionalisation of responsible innovation. *Research Policy*, 50(1): 104132, Jan. 2021. ISSN 00487333. doi: 10.1016/j.respol.2020.104132. URL https://linkinghub.elsevier.com/retrieve/pii/S0048733320302079.

Oxford Economics. How Robots Change the World, June 2019. URL https://www.oxfordeconomics.com/resource/how-robots-change-the-world/.

O. O'Neill. Linking Trust to Trustworthiness. *International Journal of Philosophical Studies*, 26(2):293–300, Mar. 2018. ISSN 0967-2559, 1466-4542. doi: 10.1080/09672559.2018.1454637. URL https://www.tandfonline.com/doi/full/10.1080/09672559.2018.1454637.

L. Pacchiardi, A. J. Chan, S. Mindermann, I. Moscovitz, A. Y. Pan, Y. Gal, O. Evans, and J. Brauner. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. *arXiv preprint arXiv:2309.15840*, 2023.

A. K. Palmer and A. Spirling. Large Language Models Can Argue in Convincing and Novel Ways About Politics: Evidence from Experiments and Human Judgement, May 2023. URL https://arthurspirling.org/documents/llm.pdf.

J. Pamment. The EU's Role in Fighting Disinformation: Crafting A Disinformation Framework, 2023. URL https://carnegieendowment.org/2020/09/24/eu-s-role-in-fighting-disinformation-crafting-disinformation-framework-pub-82720.

A. Pan, K. Bhatia, and J. Steinhardt. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models, Feb. 2022. URL http://arxiv.org/abs/2201.03544. arXiv:2201.03544 [cs, stat].

L. Pannozzo. *The 2008 Nova Scotia GPI accounts: indicators of genuine progress*. GPI Atlantic, 2009.

I. Papaioannou, A. C. Curry, J. L. Part, I. Shalyminov, X. Xu, Y. Yu, O. Dušek, V. Rieser, and O. Lemon. An Ensemble Model with Ranking for Social Dialogue, Dec. 2017. URL http://arxiv.org/abs/1712.07558. arXiv:1712.07558 [cs].

A. Paranjape, A. See, K. Kenealy, H. Li, A. Hardy, P. Qi, K. R. Sadagopan, N. M. Phu, D. Soylu, and C. D. Manning. Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations, Sept. 2020. URL http://arxiv.org/abs/2008.12348. arXiv:2008.12348 [cs].

B. Paranjape, S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro. ART: Automatic multi-step reasoning and tool-use for large language models, Mar. 2023. URL http://arxiv.org/abs/2303.09014. arXiv:2303.09014 [cs].

D. Parfit. *Reasons and persons*. OUP Oxford, 1984.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023a.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, Aug. 2023b. URL http://arxiv.org/abs/2304.03442. arXiv:2304.03442 [cs].

P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI Deception: A Survey of Examples, Risks, and Potential Solutions, Aug. 2023c. URL http://arxiv.org/abs/2308.14752. arXiv:2308.14752 [cs].

Y. Park and J. V. Chen. Acceptance and adoption of the innovative use of smartphone. *Industrial Management & Data Systems*, 107(9): 1349–1365, Nov. 2007. ISSN 0263-5577. doi: 10.1108/02635570710834009. URL https://www.emerald.com/insight/content/doi/10.1108/02635570710834009/full/html.

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon Emissions and Large Neural Network Training, Apr. 2021. URL http://arxiv.org/abs/2104.10350. arXiv:2104.10350 [cs].

D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink, Apr. 2022. URL http://arxiv.org/abs/2204.05149. arXiv:2204.05149 [cs].

A. Paul. The Risks of Generative AI in the Stock Market, 2023. URL https://www.opengrowth.com/resources/the-risks-of-generative-ai-in-the-stock-market.

L. A. Paul. *Transformative Experience*. OUP Oxford, Nov. 2014. ISBN 9780191027802. Google-Books-ID: E4XjBAAAQBAJ.

S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot, Feb. 2023. URL http://arxiv.org/abs/2302.06590. arXiv:2302.06590 [cs].

I. Pentina, T. Hancock, and T. Xie. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600, Mar. 2023. ISSN 07475632. doi: 10.1016/j.chb.2022.107600. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563222004204.

E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red Teaming Language Models with Language Models, Feb. 2022a. URL http://arxiv.org/abs/2202.03286. arXiv:2202.03286 [cs].

E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022b.

B. Perrigo. Bing's ai is threatening users. that's no laughing matter. *Time*, Feb. 2023. URL https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

C. Perrow. *Normal Accidents: Living with High Risk Technologies*. Princeton University Press, Oct. 1999. ISBN 9780691004129. URL https://press.princeton.edu/books/paperback/9780691004129/normal-accidents.

D. Peters, R. A. Calvo, and R. M. Ryan. Designing for motivation, engagement and wellbeing in digital experience. *Frontiers in Psychology*, page 797, 2018.

G. Petropoulos, D. Pichler, and F. Chiacchio. The impact of industrial robots on EU employment and wages: A local labour market approach. Technical report, Bruegel, Apr. 2018. URL https://www.bruegel.org/working-paper/impact-industrial-robots-eu-employment-and-wages-local-labour-market-approach.

Pfizer. How a Novel 'Incubation Sandbox' Helped Speed Up Data Analysis in Pfizer's COVID-19 Vaccine Trial, 2023. URL https://www.pfizer.com/news/articles/how_a_novel_incubation_sandbox_helped_speed_up_data_analysis_in_pfizer_s_covid_19_vaccine_trial.

J. Phillips and K. C. Land. The link between unemployment and crime rate fluctuations: An analysis at the county, state, and national levels. *Social Science Research*, 41(3):681–694, May 2012. ISSN 0049089X. doi: 10.1016/j.ssresearch.2012.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0049089X12000026.

M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Deletang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, and T. Shevlane. Evaluating frontier models for dangerous capabilities, 2024.

K. E. Pickett and R. G. Wilkinson. Income inequality and health: A causal review. *Social Science & Medicine*, 128:316–326, Mar. 2015. ISSN 02779536. doi: 10.1016/j.socscimed.2014.12.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0277953614008399.

Y. Pinsky. Bard can now connect to your Google apps and services, Sept. 2023. URL https://blog.google/products/bard/google-bard-new-features-update-sept-2023/.

V. Pitardi and H. R. Marriott. Alexa, she's not human but... unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4):626–642, 2021.

J. C. Pitt. It's Not About Technology. *Knowledge, Technology & Policy*, 23(3-4):445–454, Dec. 2010. ISSN 0897-1986, 1874-6314. doi: 10.1007/s12130-010-9125-5. URL http://link.springer.com/10.1007/s12130-010-9125-5.

B. Plank. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731.

M.-T. Png. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1434–1445, Seoul Republic of Korea, June 2022. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533200. URL https://dl.acm.org/doi/10.1145/3531146.3533200.

K. Polanyi. *The Great Transformation: The Political and Economic Origins of Our Time*. Farrar & Rinehart, New York, 1st edition, 1944. URL https://inctpped.ie.ufrj.br/spiderweb/pdf_4/Great_Transformation.pdf.

Z. Porter, I. Habli, J. McDermid, and M. Kaas. A Principles-based Ethics Assurance Argument Pattern for AI and Autonomous Systems. *AI and Ethics*, June 2023. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-023-00297-2. URL http://arxiv.org/abs/2203.15370. arXiv:2203.15370 [cs].

A. Poushneh. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, 58:102283, Jan. 2021. ISSN 09696989. doi: 10.1016/j.jretconser.2020.102283. URL https://linkinghub.elsevier.com/retrieve/pii/S0969698920312911.

B. Pouwels, J. Siegers, and J. D. Vlasblom. Income, working hours, and happiness. *Economics Letters*, 99(1):72–74, 2008. ISSN 0165-1765. URL https://econpapers.repec.org/article/eeeecolet/v_3a99_3ay_3a2008_3ai_3a1_3ap_3a72-74.htm.

V. Prabhakaran, M. Mitchell, T. Gebru, and I. Gabriel. A human rights-based approach to responsible ai. *arXiv preprint arXiv:2210.02667*, 2022.

R. Prescott-Allen. *The wellbeing of nations: A country-by-country index of quality of life and the environment*. Island press, 2001.

D. Proudfoot. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6):950–957, Apr. 2011. ISSN 00043702. doi: 10.1016/j.artint.2011.01.006. URL https://linkinghub.elsevier.com/retrieve/pii/S000437021100018X.

C. E. A. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, 3(2):104–110, Feb. 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00298-y. URL https://www.nature.com/articles/s42256-021-00298-y.

A. K. Przybylski and N. Weinstein. A large-scale test of the Goldilocks hypothesis: Quantifying the relations between digital-screen use and the mental well-being of adolescents. *Psychological Science*, 28(2):204–215, 2017.

J. Pugh. *Autonomy, rationality, and contemporary bioethics*. Oxford University Press, 2020.

A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor. " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 2853–2859, 2017.

A. Qammar, H. Wang, J. Ding, A. Naouri, M. Daneshmand, and H. Ning. Chatbots to chatgpt in a cybersecurity space: Evolution, vulnerabilities, attacks, challenges, and future recommendations. (arXiv:2306.09255), May 2023. URL http://arxiv.org/abs/2306.09255. arXiv:2306.09255 [cs].

Qualcomm. The future of AI is hybrid. Technical report, Qualcomm, May 2023. URL https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Whitepaper-The-future-of-AI-is-hybrid-Part-1-Unlocking-the-generative-AI-future-with-on-device-and-hybrid-AI.pdf.

Qubit Labs. How Many Programmers are there in the World and in the US?, Nov. 2022. URL https://qubit-labs.com/how-many-programmers-in-the-world/.

L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1993. ISBN 9780130151575. URL https://www.amazon.co.uk/Fundamentals-Speech-Recognition-Prentice-Processing/dp/0130151572. Google-Books-ID: XEVqQgAACAAJ.

E. Racine, T. Martin Rubio, J. Chandler, C. Forlini, and J. Lucke. The value and pitfalls of speculation about science and technology in bioethics: the case of cognitive enhancement. *Medicine, Health Care and Philosophy*, 17(3):325–337, Aug. 2014. ISSN 1386-7423, 1572-8633. doi: 10.1007/s11019-013-9539-4. URL http://link.springer.com/10.1007/s11019-013-9539-4.

A. Radhakrishnan, B. Shlegeris, R. Greenblatt, and F. Roger. Scalable oversight and weak-to-strong generalization: Compatible approaches to the same problem, 2023. URL https://www.alignmentforum.org/posts/hw2tGSsvLLyjFoLFS/scalable-oversight-and-weak-to-strong-generalization.

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. v. d. Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. d. M. d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. d. L. Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling Language Models: Methods, Analysis & Insights from Training Gopher, Jan. 2022. URL https://arxiv.org/pdf/2112.11446.pdf. arXiv:2112.11446 [cs].

M. Raghavan and S. Barocas. Challenges for mitigating bias in algorithmic hiring, Dec. 2019. URL https://www.brookings.edu/articles/challenges-for-mitigating-bias-in-algorithmic-hiring/.

I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, New York NY USA, Feb. 2020a. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375820. URL https://dl.acm.org/doi/10.1145/3375627.3375820.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 33–44, New York, NY, USA, Jan. 2020b. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022a.

I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, June 2022b. URL http://arxiv.org/abs/2206.04737. arXiv:2206.04737 [cs].

S. Ranjan. Unveiling the Privacy Risks of Generative AI, June 2023. URL https://medium.com/@shiveshr/unveiling-the-privacy-risks-of-generative-ai-d4852be407cb.

N. D. Rao and J. Min. Decent living standards: Material prerequisites for human wellbeing. *Social Indicators Research*, 138:225–244, 2018.

A. Rapoport and A. M. Chammah. The Game of Chicken. *American Behavioral Scientist*, 10(3):10–28, Nov. 1966. ISSN 0002-7642, 1552-3381. doi: 10.1177/000276426601000303. URL http://journals.sagepub.com/doi/10.1177/000276426601000303.

H. Rashkin, D. Reitter, G. S. Tomar, and D. Das. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL https://aclanthology.org/2021.acl-long.58.

P. A. Rauschnabel and A. C. Ahuvia. You're so lovable: Anthropomorphism and brand love. *Journal of Brand Management*, 21(5):372–395, June 2014. ISSN 1350-231X, 1479-1803. doi: 10.1057/bm.2014.14. URL http://link.springer.com/10.1057/bm.2014.14.

S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed. Skillful Precipitation Nowcasting using Deep Generative Models of Radar. *Nature*, 597(7878):672–677, Sept. 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03854-z. URL http://arxiv.org/abs/2104.00954. arXiv:2104.00954 [cs].

J. Raz. *Engaging reason: On the theory of value and action*. Oxford University Press, 1999.

J. Reason. *Managing the Risks of Organizational Accidents*. Routledge, Dec. 1997. ISBN 9781840141054. URL https://www.routledge.com/Managing-the-Risks-of-Organizational-Accidents/Reason/p/book/9781840141054.

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A Generalist Agent. *Transactions on Machine Learning Research*, Nov. 2022. doi: 10.48550/arXiv.2205.06175. URL https://arxiv.org/pdf/2205.06175.pdf?fs=e&s=cl. arXiv:2205.06175 [cs].

N. Reich and F. Eyssel. Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Paladyn, Journal of Behavioral Robotics*, 4(2):123–130, 2013.

M. Rheu, J. Y. Shin, W. Peng, and J. Huh-Yoo. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction*, 37(1):81–96, 2021. doi: 10.1080/10447318.2020.1807710. URL https://doi.org/10.1080/10447318.2020.1807710.

B. Ribeiro, R. Meckin, A. Balmer, and P. Shapira. The digitalisation paradox of everyday scientific labour: How mundane knowledge work is amplified and diversified in the biosciences. *Research Policy*, 52(1):104607, Jan. 2023. ISSN 00487333. doi: 10.1016/j.respol.2022.104607. URL https://linkinghub.elsevier.com/retrieve/pii/S0048733322001305.

M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira. Auditing Radicalization Pathways on YouTube, Oct. 2021. URL http://arxiv.org/abs/1908.08313. arXiv:1908.08313 [cs].

P. Ribino. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(Suppl 1):445–482, 2023.

J. C. Ribot and N. L. Peluso. A Theory of Access*. *Rural Sociology*, 68(2):153–181, June 2003. ISSN 0036-0112, 1549-0831. doi: 10.1111/j.1549-0831.2003.tb00133.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1549-0831.2003.tb00133.x.

B. Richards, B. Agüera y Arcas, G. Lajoie, and D. Sridhar. The Illusion Of AI's Existential Risk. *Noema*, July 2023. URL https://www.noemamag.com/the-illusion-of-ais-existential-risk.

R. Richardson. Defining and Demystifying Automated Decision Systems, Mar. 2021. URL https://papers.ssrn.com/abstract=3811708.

J. Richens, R. Beard, and D. H. Thompson. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.

G. Rieder, J. Simon, and P.-H. Wong. Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums, Oct. 2020. URL https://papers.ssrn.com/abstract=3717451.

V. Rieser and O. Lemon. *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Springer, Berlin, Heidelberg, 2011. ISBN 9783642249419 9783642249426. doi: 10.1007/978-3-642-24942-6. URL https://link.springer.com/10.1007/978-3-642-24942-6.

K. K. Rigaud, A. M. De Sherbinin, B. Jones, J. Bergmann, V. Clement, K. Ober, J. Schewe, S. B. Adamo, B. McCusker, S. Heuser, and A. Midgley. Groundswell : Preparing for Internal Climate Migration. 2018. doi: 10.7916/D8Z33FNS. URL https://academiccommons.columbia.edu/doi/10.7916/D8Z33FNS.

A. Rigot. Design From the Margins: Centering the most marginalized and impacted in design processes – from ideation to production. Technical report, Harvard Kennedy School: Belfer Center, May 2022. URL https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Afsaneh_Design%20From%20the%20Margins_Final_220514.pdf.

S. Rismani, R. Shelby, A. Smart, E. Jatho, J. Kroll, A. Moon, and N. Rostamzadeh. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML, Oct. 2022. URL http://arxiv.org/abs/2210.03535. arXiv:2210.03535 [cs].

H. Ritchie, M. Roser, and P. Rosado. CO_2 and Greenhouse Gas Emissions. *Our World in Data*, May 2020. URL https://ourworldindata.org/co2-and-greenhouse-gas-emissions.

H. Ritchie, V. Samborska, N. Ahuja, E. Ortiz-Ospina, and M. Roser. Global Education. *Our World in Data*, Nov. 2023. URL https://ourworldindata.org/global-education.

H. W. J. Rittel and M. M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, June 1973. ISSN 1573-0891. doi: 10.1007/BF01405730. URL https://doi.org/10.1007/BF01405730.

L. Robbins. Interpersonal comparisons of utility: A comment. *The Economic Journal*, 48(192):635–641, 1938.

D. Roberts and S. Jesudason. Movement Intersectionality: The Case of Race, Gender, Disability, and Genetic Technologies. *Du Bois Review: Social Science Research on Race*, 10(2):313–328, 2013. ISSN 1742-058X, 1742-0598. doi: 10.1017/S1742058X13000210. URL https://www.cambridge.org/core/product/identifier/S1742058X13000210/type/journal_article.

P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–108, Christchurch, New Zealand, Mar. 2016. IEEE. ISBN 9781467383707. doi: 10.1109/HRI.2016.7451740. URL http://ieeexplore.ieee.org/document/7451740/.

L. Robinson, S. R. Cotten, H. Ono, A. Quan-Haase, G. Mesch, W. Chen, J. Schulz, T. M. Hale, and M. J. Stern. Digital inequalities and why they matter. *Information, Communication & Society*, 18(5):569–582, May 2015. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2015.1012532. URL http://www.tandfonline.com/doi/abs/10.1080/1369118X.2015.1012532.

Y. M. Rocha, G. A. De Moura, G. A. Desidério, C. H. De Oliveira, F. D. Lourenço, and L. D. De Figueiredo Nicolete. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health*, 31(7):1007–1016, July 2023. ISSN 2198-1833, 1613-2238. doi: 10.1007/s10389-021-01658-z. URL https://link.springer.com/10.1007/s10389-021-01658-z.

E. Roesler, D. Manzey, and L. Onnasch. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58), Sept. 2021. ISSN 2470-9476. doi: 10.1126/scirobotics.abj5425. URL https://www.science.org/doi/10.1126/scirobotics.abj5425.

R. Rogers. Is GPT-4 Worth the Subscription? Here's What You Should Know. *Wired*, 2023. ISSN 1059-1028. URL https://www.wired.com/story/what-is-chatgpt-plus-gpt4-openai/.

M. Rojas and J. Guardiola. A hierarchy of unsatisfied needs: A subjective well-being study. *A Life Devoted to Quality of Life: Festschrift in Honor of Alex C. Michalos*, pages 105–122, 2016.

M. Rojas, A. Méndez, and K. Watkins-Fassler. The hierarchy of needs empirical examination of Maslow's theory and lessons for development. *World Development*, 165:106185, 2023.

D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio. Tackling Climate Change with Machine Learning, Nov. 2019. URL http://arxiv.org/abs/1906.05433. arXiv:1906.05433 [cs, stat].

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, Apr. 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

K. Roose. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*, Feb. 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

J. Roozenbeek, S. Van Der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254, 2022.

G. Rosen. Investments to Fight Polarization, May 2020. URL https://about.fb.com/news/2020/05/investments-to-fight-polarization/.

S. I. Ross, F. Martinez, S. Houde, M. Muller, and J. D. Weisz. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 491–514, Sydney NSW Australia, Mar. 2023. ACM. ISBN 9798400701061. doi: 10.1145/3581641.3584037. URL https://dl.acm.org/doi/10.1145/3581641.3584037.

M. Rossignac-Milon, N. Bolger, K. S. Zee, E. J. Boothby, and E. T. Higgins. Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology*, 120(4):882–911, Apr. 2021. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspi0000266. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/pspi0000266.

V. Rotondi, L. Stanca, and M. Tomasuolo. Connecting alone: Smartphone use, quality of social interactions and well-being. *Journal of Economic Psychology*, 63:17–26, Dec. 2017. ISSN 01674870. doi: 10.1016/j.joep.2017.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0167487017302520.

D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3):393–404, July 1998. ISSN 0363-7425, 1930-3807. doi: 10.5465/amr.1998.926617. URL http://journals.aom.org/doi/10.5465/amr.1998.926617.

N. Rowe. 'It's destroyed me completely': Kenyan moderators decry toll of training of AI models. *The Guardian*, Aug. 2023. ISSN 0261-3077. URL https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai.

B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code Llama: Open Foundation Models for Code, Aug. 2023. URL http://arxiv.org/abs/2308.12950. arXiv:2308.12950 [cs].

G. Rubeis. The disruptive power of Artificial Intelligence. Ethical aspects of gerontechnology in elderly care. *Archives of Gerontology and Geriatrics*, 91:104186, Nov. 2020. ISSN 01674943. doi: 10.1016/j.archger.2020.104186. URL https://linkinghub.elsevier.com/retrieve/pii/S0167494320301801.

A. Rubel, A. Pham, and C. Castro. Agency Laundering and Algorithmic Decision Systems. In N. Taylor, C. Christian-Lamb, M. Martin, and B. Nardi, editors, *Information in Contemporary Society (Lecture Notes in Computer Science) (Proceedings of the 2019 iConference)*, pages 590–598. Springer Nature, 2019.

S. Russell and P. Norvig. *Artificial intelligence a modern approach*. Prentice Hall, 1995.

S. J. Russell. *Human compatible: artificial intelligence and the problem of control*. Viking, New York?, 2019. ISBN 9780525558620.

M. Ryan. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5):2749–2767, Oct. 2020. ISSN 1471-5546. doi: 10.1007/s11948-020-00228-y. URL https://doi.org/10.1007/s11948-020-00228-y.

R. M. Ryan, R. R. Curren, and E. L. Deci. What humans need: Flourishing in aristotelian philosophy and self-determination theory. In A. S. Waterman, editor, *The best within us: Positive psychology perspectives on eudaimonia*, pages 57–75. American Psychological Association, 2013.

H. Ryland. It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships. *Minds and Machines*, 31(3):377–393, Sept. 2021. ISSN 1572-8641. doi: 10.1007/s11023-021-09560-z. URL https://doi.org/10.1007/s11023-021-09560-z.

C. Rzepka, B. Berger, and T. Hess. Voice Assistant vs. Chatbot – Examining the Fit Between Conversational Agents' Interaction Modalities and Information Search Tasks. *Information Systems Frontiers*, 24(3):839–856, June 2022. ISSN 1572-9419. doi: 10.1007/s10796-021-10226-5. URL https://doi.org/10.1007/s10796-021-10226-5.

T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks, Aug. 2023. URL http://arxiv.org/abs/2207.13243. arXiv:2207.13243 [cs].

V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-Generated Text be Reliably Detected?, June 2023. URL http://arxiv.org/abs/2303.11156. arXiv:2303.11156 [cs].

R. S. Sai Dinesh, R. Surendran, D. Kathirvelan, and V. Logesh. Artificial Intelligence based Vision and Voice Assistant. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1478–1483, Tuticorin, India, Mar. 2022. IEEE. ISBN 9781665484251. doi: 10.1109/ICEARS53579.2022.9751819. URL https://ieeexplore.ieee.org/document/9751819/.

R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir. Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education, Sept. 2023. URL http://arxiv.org/abs/2309.10892. arXiv:2309.10892 [cs].

M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5:313–323, 2013.

S. Salles. Theories of Vagueness. In S. Salles, editor, *Vagueness as Arbitrariness: Outline of a Theory of Vagueness*, Synthese Library, pages 65–128. Springer International Publishing, Cham, 2021. ISBN 9783030667818. doi: 10.1007/978-3-030-66781-8_4. URL https://doi.org/10.1007/978-3-030-66781-8_4.

N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada, Mar. 2021. ACM. ISBN 9781450383097. doi: 10.1145/3442188.3445896. URL https://dl.acm.org/doi/10.1145/3442188.3445896.

M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.pdf.

R. Sarikaya, P. A. Crook, A. Marin, M. Jeong, J. Robichaud, A. Celikyilmaz, Y. Kim, A. Rochette, O. Z. Khan, X. Liu, D. Boies, T. Anastasakos, Z. Feizollahi, N. Ramesh, H. Suzuki, R. Holenstein, E. Krawczyk, and V. Radostev. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 391–397, San Diego, CA, Dec. 2016. IEEE. ISBN 9781509049035. doi: 10.1109/SLT.2016.7846294. URL http://ieeexplore.ieee.org/document/7846294/.

G. Sartor, F. Lagioia, and F. Galli. Regulating targeted and behavioural advertising in digital services. How to ensure users' informed consent | Think Tank | European Parliament. Technical Report PE 694.680, European Parliament's Committee on Legal Affairs, Sept. 2021. URL https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2021)694680.

L. Sartori and A. Theodorou. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1):4, Jan. 2022. ISSN 1572-8439. doi: 10.1007/s10676-022-09624-3. URL https://doi.org/10.1007/s10676-022-09624-3.

D. Satz. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets*. Oxford University Press, June 2010. ISBN 9780199718573. Google-Books-ID: h13Pk15YpIwC.

L. Saulnier, S. Karamcheti, H. Laurençon, L. Tronchon, T. Wang, V. Sanh, A. Singh, G. Pistilli, S. Luccioni, Y. Jernite, M. Mitchell, and D. Kiela. Putting ethical principles at the core of the research lifecycle, May 2022. URL https://huggingface.co/blog/ethical-charter-multimodal.

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators, June 2022. URL https://arxiv.org/pdf/2206.05802.pdf. arXiv:2206.05802 [cs].

T. Scanlon. *What We Owe to Each Other*. Belknap Press of Harvard University Press, Cambridge, Mass., 1998.

R. Schaeffer, B. Miranda, and S. Koyejo. Are Emergent Abilities of Large Language Models a Mirage?, May 2023. URL https://arxiv.org/pdf/2304.15004.pdf. arXiv:2304.15004 [cs].

J. Schaubroeck, S. S. K. Lam, and A. C. Peng. Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. *Journal of Applied Psychology*, 96(4):863–871, 2011. ISSN 1939-1854, 0021-9010. doi: 10.1037/a0022625. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0022625.

J. Scheurer, J. A. Campos, J. S. Chan, A. Chen, K. Cho, and E. Perez. Training Language Models with Language Feedback, Nov. 2022. URL http://arxiv.org/abs/2204.14146. arXiv:2204.14146 [cs].

M. Scheutz. The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. Jan. 2009.

T. Schick, S. Udupa, and H. Schütze. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, Dec. 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00434. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00434/108865/Self-Diagnosis-and-Self-Debiasing-A-Proposal-for.

T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, Feb. 2023. URL https://arxiv.org/pdf/2302.04761.pdf. arXiv:2302.04761 [cs].

J.-E. Schirmer. Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany. In T. Wischmeyer and T. Rademacher, editors, *Regulating Artificial Intelligence*, pages 123–142. Springer International Publishing, Cham, 2020. ISBN 9783030323608 9783030323615. doi: 10.1007/978-3-030-32361-5_6. URL https://link.springer.com/10.1007/978-3-030-32361-5_6.

D. Schlangen. Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research, Aug. 2019. URL http://arxiv.org/abs/1908.10747. arXiv:1908.10747 [cs].

D. Schlangen and G. Skantze. A General, Abstract Model of Incremental Dialogue Processing. In A. Lascarides, C. Gardent, and J. Nivre, editors, *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece, Mar. 2009. Association for Computational Linguistics. URL https://aclanthology.org/E09-1081.

J. Schroeder and N. Epley. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology. General*, 145(11):1427–1437, Nov. 2016. ISSN 1939-2222. doi: 10.1037/xge0000214.

J. M. Schröder and M. Neumayr. How socio-economic inequality affects individuals' civic engagement: a systematic literature review of empirical findings and theoretical explanations. *Socio-Economic Review*, 21(1):665–694, Mar. 2023. ISSN 1475-1461, 1475-147X. doi: 10.1093/ser/mwab058. URL https://academic.oup.com/ser/article/21/1/665/6482042.

P. M. Schwartz and D. J. Solove. The pii problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, 86:1814, 2011.

R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, Nov. 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3381831. URL https://dl.acm.org/doi/10.1145/3381831.

M. Schwerin. Somehow, Airline Customer Service Is Getting Even Worse, June 2023. URL https://www.theatlantic.com/technology/archive/2023/06/airline-customer-service-chatbot-ai/674412/.

M. S. Schäfer. The Notorious GPT: science communication in the age of artificial intelligence. *Journal of Science Communication*, 22(02), May 2023. ISSN 1824-2049. doi: 10.22323/2.22020402. URL https://jcom.sissa.it/article/pubid/JCOM_2202_2023_Y02/.

E. Seger, S. Avin, G. Pearson, M. Briers, S. Ó Heigeartaigh, and H. Bacon. Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. Technical report, The Alan Turing Institute, Oct. 2020. URL https://www.repository.cam.ac.uk/handle/1810/317073.

E. Seger, A. Ovadya, D. Siddarth, B. Garfinkel, and A. Dafoe. Democratising AI: Multiple Meanings, Goals, and Methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 715–722, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604693. URL https://dl.acm.org/doi/10.1145/3600211.3604693.

A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, Jan. 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

A. Sen. *Development as freedom*. Oxford Paperbacks, 2001.

J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute Trends Across Three Eras of Machine Learning, Mar. 2022. URL http://arxiv.org/abs/2202.05924. arXiv:2202.05924 [cs].

W. Seymour, X. Zhan, M. Cote, and J. Such. A Systematic Review of Ethical Concerns with Voice Assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–145, Aug. 2023. doi: 10.1145/3600211.3604679. URL http://arxiv.org/abs/2211.04193. arXiv:2211.04193 [cs].

H. Shah. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2128):20170362, 2018.

R. Shah. AI Risk from Program Search in Threat Model Literature Review, 2022. URL https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review.

R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals, Nov. 2022. URL http://arxiv.org/abs/2210.01790. arXiv:2210.01790 [cs].

C. Shaib, M. L. Li, S. Joseph, I. J. Marshall, J. J. Li, and B. C. Wallace. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success), May 2023. URL http://arxiv.org/abs/2305.06299. arXiv:2305.06299 [cs].

M. Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024. URL http://dl.acm.org/doi/10.1145/3624724.

M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, pages 1–6, 2023.

S. Shankland. Photoshop's Firefly Generative AI Arrives With a Creative Cloud Price Hike. *CNET*, Sept. 2023. URL https://www.cnet.com/tech/computing/photoshops-firefly-generative-ai-arrives-with-a-creative-cloud-price-hike/.

M. Shardlow and P. Przybyła. Deanthropomorphising NLP: Can a Language Model Be Conscious?, Nov. 2023. URL http://arxiv.org/abs/2211.11483. arXiv:2211.11483 [cs].

Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, K. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D. G. Robinson. Practices for governing agentic AI systems. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf, 2023. Accessed: 2023-01-04.

J. Shaw. Content moderators pay a psychological toll to keep social media clean. We should be helping them. https://www.sciencefocus.com/news/content-moderators-pay-a-psychological-toll-to-keep-social-media-clean-we-should-be-helping-them, 2022. Accessed: 2023-01-04.

M. Sheehan, P. Friesen, A. Balmer, C. Cheeks, S. Davidson, J. Devereux, D. Findlay, K. Keats-Rohan, R. Lawrence, and K. Shafiq. Trust, trustworthiness and sharing patient data for research. *Journal of Medical Ethics*, 47(12):26–26, 2021. doi: 10.1136/medethics-2019-106048.

R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, and G. Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 9798400702310. doi: 10.1145/3600211.3604673. URL https://dl.acm.org/doi/10.1145/3600211.3604673.

X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models, Aug. 2023. URL https://arxiv.org/pdf/2308.03825.pdf. arXiv:2308.03825 [cs].

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe. Model evaluation for extreme risks, Sept. 2023. URL http://arxiv.org/abs/2305.15324. arXiv:2305.15324 [cs].

H. Shevlin and M. Halina. Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, 1(4):165–167, 2019.

R. Shiffrin and M. Mitchell. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120, Mar. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2300963120. URL https://pnas.org/doi/10.1073/pnas.2300963120.

S. V. Shiffrin. Paternalism, unconscionability doctrine, and accommodation. *Philosophy & Public Affairs*, 29(3):205–250, 2000.

H. I. Shin and J. Kim. My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology*, 39(2):445–453, Apr. 2020. ISSN 1046-1310, 1936-4733. doi: 10.1007/s12144-018-9975-7. URL http://link.springer.com/10.1007/s12144-018-9975-7.

V. K. M. Shiramizu, A. J. Lee, D. Altenburg, D. R. Feinberg, and B. C. Jones. The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports*, 12(1):22479, Dec. 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-27124-8. URL https://www.nature.com/articles/s41598-022-27124-8.

L. Siddharth, L. Blessing, and J. Luo. Natural language processing in-and-for design research. *Design Science*, 8:e21, 2022. ISSN 2053-4701. doi: 10.1017/dsj.2022.16. URL https://www.cambridge.org/core/product/identifier/S2053470122000166/type/journal_article.

D. Siegel and M. Bennett Doty. Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda, Feb. 2023. URL https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.

D. Siemon, T. Strohmann, B. Khosrawi-Rad, T. d. Vreede, E. Elshan, and M. Meyer. Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. *AMCIS 2022 Proceedings*, Aug. 2022. URL https://aisel.aisnet.org/amcis2022/sig_hci/sig_hci/10.

A. J. Simmons. Justification and Legitimacy. *Ethics*, 109(4):739–771, 1999. doi: 10.1086/233944.

J. Sims. BlackMamba: Using AI to Generate Polymorphic Malware, Mar. 2023. URL https://securityboulevard.com/2023/03/blackmamba-using-ai-to-generate-polymorphic-malware/.

R. Singel. Filtering Out the Bots: What Americans Actually Told the FCC about Net Neutrality Repeal. Technical report, Stanford Law School: The Center for Internet and Society, Oct. 2018. URL https://cyberlaw.stanford.edu/sites/default/files/FilteringOutTheBotsUniqueNetNeutralityComments.pdf.

N. Singer. Applying to College? Here's How A.I. Tools Might Hurt, or Help. *The New York Times*, Sept. 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/09/01/technology/ai-chatbots-college-applications.html.

P. Singer and Y. F. Tse. AI ethics: the case for including animals. *AI and Ethics*, 3(2):539–551, May 2023. ISSN 2730-5961. doi: 10.1007/s43681-022-00187-z. URL https://doi.org/10.1007/s43681-022-00187-z.

U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data, Sept. 2022. URL http://arxiv.org/abs/2209.14792. arXiv:2209.14792 [cs].

S. Singh-Kurtz. The Man of Your Dreams, Mar. 2023. URL https://www.thecut.com/article/ai-artificial-intelligence-chatbot-replika-boyfriend.html.

N. Sirlin, Z. Epstein, A. A. Arechar, and D. G. Rand. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review*, Dec. 2021. doi: 10.37016/mr-2020-83. URL https://misinformereview.hks.harvard.edu/article/digital-literacy-is-associated-with-more-discerning-accuracy-judgments-but-not-sharing-intentions/.

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and Characterizing Reward Hacking, Sept. 2022. URL http://arxiv.org/abs/2209.13085. arXiv:2209.13085 [cs, stat].

S. Skinner-Thompson. *Privacy at the Margins*. Cambridge University Press, 1 edition, Nov. 2020. ISBN 9781316850350 9781107181373 9781316632635. doi: 10.1017/9781316850350. URL https://www.cambridge.org/core/product/identifier/9781316850350/type/book.

M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601, 2021.

M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168, Dec. 2022. ISSN 10715819. doi: 10.1016/j.ijhcs.2022.102903. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581922001252.

Slack. Claude, 2023. URL https://slack.com/apps/A04KGS7N9A8-claude.

M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, Arlington VA USA, Oct. 2022. ACM. ISBN 9781450394772. doi: 10.1145/3551624.3555285. URL https://dl.acm.org/doi/10.1145/3551624.3555285.

C. T. Small, I. Vendrov, E. Durmus, H. Homaei, E. Barry, J. Cornebise, T. Suzman, D. Ganguli, and C. Megill. Opportunities and Risks of LLMs for Scalable Deliberation with Polis, June 2023. URL http://arxiv.org/abs/2306.11932. arXiv:2306.11932 [cs].

S. Smit, T. Tacke, S. Lund, J. Manyika, and L. Thiel. The future of work in Europe. Technical report, McKinsey Global Institute, June 2020. URL https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/the%20future%20of%20work%20in%20europe/mgi-the-future-of-work-in-europe-discussion-paper.pdf.

C. L. Smith and P. B. Kantor. User adaptation: Good results from poor systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 147–154, 2008.

H. Smith, A. Manzini, M.-R. Kennedy, and J. Ives. Ethics of Trust/worthiness in Autonomous Systems: a scoping review. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pages 1–15, Edinburgh United Kingdom, July 2023. ACM. ISBN 9798400707346. doi: 10.1145/3597512.3600207. URL https://dl.acm.org/doi/10.1145/3597512.3600207.

H. J. Smith, T. Dinev, and H. Xu. Information Privacy Research: An Interdisciplinary Review. *MIS Quarterly*, 35(4):989, 2011. ISSN 02767783. doi: 10.2307/41409970. URL https://www.jstor.org/stable/10.2307/41409970.

L. M. Smith, J. L. Case, H. M. Smith, L. C. Harwell, and J. Summers. Relating ecosystem services to domains of human well-being: Foundation for a us index. *Ecological Indicators*, 28:79–90, 2013.

A. Snyder. AI's language gap. *Axios*, Sept. 2023. URL https://www.axios.com/2023/09/08/ai-language-gap-chatgpt.

I. Solaiman. The Gradient of Generative AI Release: Methods and Considerations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 111–122, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3593981. URL https://dl.acm.org/doi/10.1145/3593013.3593981.

I. Solaiman, Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. L. Blodgett, H. Daumé III, J. Dodge, E. Evans, S. Hooker, Y. Jernite, A. S. Luccioni, A. Lusoli, M. Mitchell, J. Newman, M.-T. Png, A. Strait, and A. Vassilev. Evaluating the Social Impact of Generative AI Systems in Systems and Society, June 2023. URL http://arxiv.org/abs/2306.05949. arXiv:2306.05949 [cs].

R. Sorensen. Vagueness. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2023 edition, 2023. URL https://plato.stanford.edu/archives/win2023/entries/vagueness/.

G. Spadaro, K. Gangl, J.-W. Van Prooijen, P. A. Van Lange, and C. O. Mosso. Enhancing feelings of security: How institutional trust promotes interpersonal trust. *PloS one*, 15(9):e0237934, 2020.

G. Spitale, N. Biller-Andorno, and F. Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850, June 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adh1850. URL https://www.science.org/doi/10.1126/sciadv.adh1850.

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson, C. Argueta, C. F. Ramírez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodola, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. Wang, G. Jaimovitch-López, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocoń, J. Thompson, J. Wingfield, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. Omondi, K. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Colón, L. Metz, L. K. Şenel, M. Bosma, M. Sap, M. ter Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. R. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts, N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. LeBras, R. Liu,

R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, Shyamolima, Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, and Z. Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2023. URL http://arxiv.org/abs/2206.04615. arXiv:2206.04615 [cs, stat].

Stack Overflow. Stack Overflow Developer Survey 2023, 2023. URL https://survey.stackoverflow.co/2023/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2023.

H. v. Stackelberg. *Marktform und Gleichgewicht*. J. Springer, 1934. Google-Books-ID: wihBAAAAIAAJ.

G. Starke and M. Ienca. Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics*, pages 1–10, Oct. 2022. ISSN 0963-1801, 1469-2147. doi: 10.1017/S0963180122000445. URL https://www.cambridge.org/core/product/identifier/S0963180122000445/type/journal_article.

R. Starkman. Stochastic Parrots and Shaky Foundations, Aug. 2021. URL https://ruth.substack.com/p/stochastic-parrots-and-shaky-foundations.

State of California. Bots: disclosure, Sept. 2018. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001.

P. Steca, D. Monzani, A. Greco, M. D'Addario, E. Cappelletti, and L. Pancani. The effects of short-term personal goals on subjective well-being. *Journal of Happiness Studies*, 17:1435–1450, 2016.

N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback, Feb. 2022. URL http://arxiv.org/abs/2009.01325. arXiv:2009.01325 [cs].

J. Stilgoe, R. Owen, and P. Macnaghten. Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580, Nov. 2013. ISSN 00487333. doi: 10.1016/j.respol.2013.05.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0048733313000930.

C. Stokel-Walker. The Generative AI Race Has a Dirty Secret. *Wired UK*, Feb. 2023. ISSN 1357-0978. URL https://www.wired.co.uk/article/the-generative-ai-search-race-has-a-dirty-secret.

C. L. v. Straten, J. Peter, R. Kühne, and A. Barco. Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2):1–22, 2020.

P. F. Strawson. *Analysis and metaphysics: An introduction to philosophy*. Oxford University Press, USA, 1992.

J. Stray. Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, 3(4):443–463, 2020.

J. Stray, I. Vendrov, J. Nixon, S. Adler, and D. Hadfield-Menell. What are you optimizing for? Aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

J. Stray, A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, C. Bakalar, L. Beattie, M. Ekstrand, C. Leibowicz, et al. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2022.

D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with Humans without Human Data. In *Advances in Neural Information Processing Systems*, volume 34, pages 14502–14515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/797134c3e42371bb4979a462eb2f042a-Abstract.html.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i09.7123. URL https://ojs.aaai.org/index.php/AAAI/article/view/7123.

A. Strudler. Deception Unraveled. *The Journal of Philosophy*, 102(9):458–473, 2005. ISSN 0022-362X. URL https://www.jstor.org/stable/3655633.

E. K. Suckiel. William James. In J. R. Shook and J. Margolis, editors, *A Companion to Pragmatism*, pages 30–43. Blackwell Publishing Ltd, Oxford, UK, Jan. 2006. ISBN 9780470997079 9781405116213. doi: 10.1002/9780470997079.ch3. URL https://onlinelibrary.wiley.com/doi/10.1002/9780470997079.ch3.

J. L. Sullivan and J. E. Transue. The Psychological Underpinnings of Democracy: A Selective Review of Research on Political Tolerance, Interpersonal Trust, and Social Capital. *Annual Review of Psychology*, 50(1):625–650, Feb. 1999. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.psych.50.1.625. URL https://www.annualreviews.org/doi/10.1146/annurev.psych.50.1.625.

J. K. Summers, L. M. Smith, L. C. Harwell, J. L. Case, C. M. Wade, K. R. Straub, and H. M. Smith. An index of human well-being for the us: A TRIO approach. *Sustainability*, 6(6):3915–3935, 2014.

D. Summers-Stay, C. R. Voss, and S. M. Lukin. Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023. URL https://openreview.net/forum?id=8HwKaJ1wvl.

S. S. Sundar, M. D. Molina, and E. Cho. Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? *Journal of Computer-Mediated Communication*, 26(6):301–319, Nov. 2021. ISSN 1083-6101. doi: 10.1093/jcmc/zmab010. URL https://academic.oup.com/jcmc/article/26/6/301/6336055.

C. R. Sunstein. Nudges Do Not Undermine Human Agency. *Journal of Consumer Policy*, 38(3):207–210, Sept. 2015. ISSN 1573-0700. doi: 10.1007/s10603-015-9289-1. URL https://doi.org/10.1007/s10603-015-9289-1.

C. R. Sunstein. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge University Press, Aug. 2016. ISBN 9781107140707. Google-Books-ID: TlvWDAAAQBAJ.

H. Suresh and J. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA, Oct. 2021. ACM. ISBN 9781450385534. doi: 10.1145/3465416.3483305. URL https://dl.acm.org/doi/10.1145/3465416.3483305.

H. Suresh, R. Movva, A. L. Dogan, R. Bhargava, I. Cruxen, A. M. Cuba, G. Taurino, W. So, and C. D'Ignazio. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 667–678, Seoul Republic of Korea, June 2022. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533132. URL https://dl.acm.org/doi/10.1145/3531146.3533132.

D. Susser, B. Roessler, and H. Nissenbaum. Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review*, 4(1):2–45, 2019a. URL https://philarchive.org/archive/SUSOMH.

D. Susser, B. Roessler, and H. Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), June 2019b. ISSN 2197-6775. doi: 10.14763/2019.2.1410. URL https://policyreview.info/node/1410.

R. Sutton. The Bitter Lesson, Mar. 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

L. Sweeney. Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, Mar. 2013. ISSN 1542-7730, 1542-7749. doi: 10.1145/2460276.2460278. URL https://dl.acm.org/doi/10.1145/2460276.2460278.

Swiss Re Institute. The economics of climate change: no action not an option. Technical report, Swiss Re Institute, Apr. 2021. URL https://www.swissre.com/dam/jcr:e73ee7c3-7f83-4c17-a2b8-8ef23a8d3312/swiss-re-institute-expertise-publication-economics-of-climate-change.pdf.

M. Syed. *Black Box Thinking: Why Most People Never Learn from Their Mistakes–But Some Do*. Penguin, Nov. 2015. ISBN 9781591848226. Google-Books-ID: MrJPEAAAQBAJ.

H. S. Sætra. The Parasitic Nature of Social AI: Sharing Minds with the Mindless. *Integrative Psychological and Behavioral Science*, 54(2):308–326, June 2020. ISSN 1936-3567. doi: 10.1007/s12124-020-09523-6. URL https://doi.org/10.1007/s12124-020-09523-6.

M. Tabachnyk and S. Nikolov. ML-Enhanced Code Completion Improves Developer Productivity, July 2022. URL https://blog.research.google/2022/07/ml-enhanced-code-completion-improves.html.

E. Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report 100-1, National Institute of Standards and Technology (NIST), Gaithersburg, MD, Jan. 2023. URL http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton. A taxonomy and terminology of adversarial machine learning. preprint, Oct. 2019. URL https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf.

M. Tahaei, M. Constantinides, D. Quercia, S. Kennedy, M. Muller, S. Stumpf, Q. V. Liao, R. Baeza-Yates, L. Aroyo, J. Holbrook, E. Luger, M. Madaio, I. G. Blumenfeld, M. De-Arteaga, J. Vitak, and A. Olteanu. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, Hamburg Germany, Apr. 2023. ACM. ISBN 9781450394222. doi: 10.1145/3544549.3583178. URL https://dl.acm.org/doi/10.1145/3544549.3583178.

Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani. VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop, Feb. 2018. URL http://arxiv.org/abs/1707.06588. arXiv:1707.06588 [cs].

H. Tameez. YouTube's algorithm is pushing climate misinformation videos, and their creators are profiting from it, Jan. 2020. URL https://www.niemanlab.org/2020/01/youtubes-algorithm-is-pushing-climate-misinformation-videos-and-their-creators-are-profiting-from-it/.

A. Tamkin, M. Brundage, J. Clark, and D. Ganguli. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models, Feb. 2021. URL http://arxiv.org/abs/2102.02503. arXiv:2102.02503 [cs].

I. F. Tapu and T. K. Fa'agau. A New Age Indigenous Instrument: Artificial Intelligence & Its Potential for (De)colonialized Data. *Harvard Civil Rights–Civil Liberties Law Review*, 57(2), 2022. URL https://journals.law.harvard.edu/crcl/wp-content/uploads/sites/80/2023/01/ANewAgeIndigenousInstrument.pdf.

M. Tatar, J. M. Shoorekchali, M. R. Faraji, M. A. Seyyedkolaee, J. A. Pagán, and F. A. Wilson. COVID-19 vaccine inequality: A global perspective. *Journal of Global Health*, 12:03072, Oct. 2022. ISSN 2047-2978, 2047-2986. doi: 10.7189/jogh.12.03072. URL https://jogh.org/2022/jogh-12-03072.

L. Tay, M. Zyphur, and C. Batz-Barbarich. Income and Subjective Well-Being: Review, Synthesis, and Future Research. In *e-Handbook of Subjective Well-Being*. Dec. 2017.

L. Tay, J. O. Pawelski, and M. G. Keith. The role of the arts and humanities in human flourishing: A conceptual model. *The Journal of Positive Psychology*, 13(3):215–225, 2018.

M. Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, Nov. 2004. ISSN 08998256. doi: 10.1016/j.geb.2004.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0899825604000314.

R. H. Thaler and C. R. Sunstein. *Nudge: The Final Edition*. Yale University Press, 2021. ISBN 9780300262285. Google-Books-ID: Wf1AEAAAQBAJ.

S. Thanawala. AI facial recognition tech leads to wave of lawsuits from Black plaintiffs after mistaken identities end in arrests, 2023. URL https://fortune.com/2023/09/25/ai-facial-recognition-tech-lawsuits-black-plaintiffs-mistaken-identities-arrests/.

The Adaptive Agents Group. The Shibboleth Rule for Artificial Agents, Aug. 2021. URL https://hai.stanford.edu/news/shibboleth-rule-artificial-agents. publisher: Stanford University.

The Collective Intelligence Project. Whitepaper, 2023. URL https://cip.org/whitepaper.

The Cybersecurity and Infrastructure Security Agency. Tactics of Disinformation, 2022. URL https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf.

The Economist Intelligence Unit. New Schools of Thought, 2020. URL https://www.qf.org.qa/eiu.

The White House. Blueprint for an AI Bill of Rights, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

The White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, July 2023a. URL https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, 2023b.

The World Bank. Lifting 800 Million People Out of Poverty – New Report Looks at Lessons from China's Experience, Apr. 2022. URL https://www.worldbank.org/en/news/press-release/2022/04/01/lifting-800-million-people-out-of-poverty-new-report-looks-at-lessons-from-china-s-experience.

A. Theben, L. Gunderson, L. López-Fóres, G. Misuraca, and F. Lupiánez-Villanueva. Challenges and limits of an open source approach to Artificial Intelligence. Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies PE 662.908, European Parliament, May 2021. URL https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662908/IPOL_STU(2021)662908_EN.pdf.

S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, June 2021. ISSN 1422-8890. doi: 10.1007/s12525-020-00441-4. URL https://doi.org/10.1007/s12525-020-00441-4.

A. Thiel. Biometric identification technologies and the Ghanaian 'data revolution'. *The Journal of Modern African Studies*, 58(1):115–136, Mar. 2020. ISSN 0022-278X, 1469-7777. doi: 10.1017/S0022278X19000600. URL https://www.cambridge.org/core/product/identifier/S0022278X19000600/type/journal_article.

D. Thiel, M. Stroebel, and R. Portnoff. Generative ML and CSAM: Implications and Mitigations. Technical report, Stanford University: Freeman Spogli Institute, June 2023. URL https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations.

R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. LaMDA: Language Models for Dialog Applications, Feb. 2022. URL https://arxiv.org/pdf/2201.08239.pdf. arXiv:2201.08239 [cs].

J. Tien, J. Z.-Y. He, Z. Erickson, A. D. Dragan, and D. S. Brown. Causal confusion and reward misidentification in preference-based reward learning, 2023.

N. Tiku. The Google engineer who thinks the company's AI has come to life. *Washington Post*, June 2022. URL https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

T. Titchkosky. *The question of access: disability, space, meaning*. University of Toronto Press, Toronto, 2011. ISBN 9781442640269 9781442685222 9781442610002. OCLC: ocn712851646.

R. Y. Toledo, A. A. Alzahrani, and L. Martinez. A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7:96695–96711, 2019.

S. Tolmeijer, N. Zierau, A. Janson, J. S. Wahdatehagh, J. M. M. Leimeister, and A. Bernstein. Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.

C. Toups, R. Bommasani, K. A. Creel, S. H. Bana, D. Jurafsky, and P. Liang. Ecosystem-level Analysis of Deployed Machine Learning Reveals Homogeneous Outcomes, July 2023. URL http://arxiv.org/abs/2307.05862. arXiv:2307.05862 [cs].

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan,

M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of the 25th USENIX Security Symposium*, pages 601–618, Austin, Texas, 2016. URL https://www.usenix.org/sites/default/files/sec16_full_proceedings.pdf.

R. J. Trappey and A. G. Woodside. *Brand Choice: Revealing Customers' Unconscious-Automatic and Strategic Thinking Processes*. Palgrave Macmillan UK, London, 2005. ISBN 9781349523573 9780230514201. doi: 10.1057/9780230514201. URL http://link.springer.com/10.1057/9780230514201.

A. Trask, E. Bluemke, B. Garfinkel, C. G. Cuervas-Mons, and A. Dafoe. Beyond Privacy Trade-offs with Structured Transparency, Dec. 2020. URL http://arxiv.org/abs/2012.08347. arXiv:2012.08347 [cs].

J. C. Tronto. *Moral Boundaries: A Political Argument for an Ethic of Care*. Routledge, 1 edition, July 2020. ISBN 9781003070672. doi: 10.4324/9781003070672. URL https://www.taylorfrancis.com/books/9781000107777.

J. C. Tronto and B. Fisher. Toward a Feminist Theory of Caring. In E. Abel and M. Nelson, editors, *Circles of Care*, pages 36–54. SUNY Press, Albany, NY, 1990.

G. Tsai. Rational Persuasion as Paternalism. *Philosophy & Public Affairs*, 42(1):78–112, Jan. 2014. ISSN 0048-3915, 1088-4963. doi: 10.1111/papa.12026. URL https://onlinelibrary.wiley.com/doi/10.1111/papa.12026.

X. Tu, J. Zou, W. J. Su, and L. Zhang. What Should Data Science Education Do with Large Language Models?, July 2023. URL http://arxiv.org/abs/2307.02792. arXiv:2307.02792 [cs].

B. Tucker. Technocapitalist Disability Rhetoric: When Technology is Confused with Social Justice | enculturation, Apr. 2017. URL https://enculturation.net/technocapitalist-disability-rhetoric.

Z. Tufekci. Opinion | YouTube, the Great Radicalizer. *The New York Times*, Mar. 2018. ISSN 0362-4331. URL https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

S. Turkle. Authenticity in the Age of Digital Companions. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systemsinteraction Studies / Social Behaviour and Communication in Biological and Artificial Systemsinteraction Studies*, 8(3):501–517, 2007. doi: 10.1075/is.8.3.11tur.

S. Turkle. There Will Never Be an Age of Artificial Intimacy. *The New York Times*, Aug. 2018. ISSN 0362-4331. URL https://www.nytimes.com/2018/08/11/opinion/there-will-never-be-an-age-of-artificial-intimacy.html.

A. M. Turner and P. Tadepalli. Parametrically Retargetable Decision-Makers Tend To Seek Power, Oct. 2022. URL http://arxiv.org/abs/2206.13477. arXiv:2206.13477 [cs].

A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli. Optimal Policies Tend to Seek Power, Jan. 2023. URL http://arxiv.org/abs/1912.01683. arXiv:1912.01683 [cs].

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, May 2023. URL http://arxiv.org/abs/2305.04388. arXiv:2305.04388 [cs].

C. Tyler, K. L. Akerlof, A. Allegra, Z. Arnold, H. Canino, M. A. Doornenbal, J. A. Goldstein, D. Budtz Pedersen, and W. J. Sutherland. AI tools as science policy advisers? The potential and the pitfalls. *Nature*, 622(7981):27–30, Oct. 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/d41586-023-02999-3. URL https://www.nature.com/articles/d41586-023-02999-3.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process- and outcome-based feedback, Nov. 2022. URL http://arxiv.org/abs/2211.14275. arXiv:2211.14275 [cs].

UK Department for Science, Innovation and Technology. A pro-innovation approach to AI regulation. Technical report, UK Government, Mar. 2023. URL https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. OCLC: 1382788551.

UK Government. National Tutoring Programme, July 2023. URL https://explore-education-statistics.service.gov.uk/find-statistics/national-tutoring-programme.

T. Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, Mar. 2023. URL http://arxiv.org/abs/2302.08399. arXiv:2302.08399 [cs].

A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, Dec. 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752. URL https://www.jair.org/index.php/jair/article/view/12752.

UN. Inequality in Asia and the Pacific in the era of the 2030 agenda for sustainable development, 2018. URL https://www.unescap.org/publications/inequality-asia-and-pacific-era-2030-agenda-sustainable-development.

UN. Addressing the Digital Divide. Technical report, United Nations Human Settlements Programme, 2021. URL https://unhabitat.org/sites/default/files/2021/11/addressing_the_digital_divide.pdf.

UNDP. Concept and measurement of human development. *Human Development Report 1990*, 1990.

UNESCO. Transforming education from within: current trends in the status and development of teachers. Technical report, UNESCO, 2022. URL https://unesdoc.unesco.org/ark:/48223/pf0000383002.

UNESCO. Transforming education together: the Global Education Coalition in action. Technical report, UNESCO, 2023. URL https://unesdoc.unesco.org/ark:/48223/pf0000384812.

UNFCCC. Secretariat. Technical dialogue of the first global stocktake. synthesis report by the co-facilitators on the technical dialogue. Sept. 2023.

UNICEF. Education, June 2022. URL https://data.unicef.org/topic/gender/gender-disparities-in-education/.

Université de Montréal. The Montréal Declaration for a Responsible Development of Artificial Intelligence, 2017. URL https://montrealdeclaration-responsibleai.com/the-declaration/.

K. Ura. Explanation of GNH index. Gross National Happiness. The Centre for Bhutan Studies, 2008.

US Bureau of Labor Statistics. The American Time Use Survey (ATUS). URL https://www.bls.gov/tus/.

US Department of Justice. Court Issues Order Requiring Cigarette Companies to Post Corrective Statements; Resolves Historic RICO Tobacco Litigation, Dec. 2022. URL https://www.justice.gov/opa/pr/court-issues-order-requiring-cigarette-companies-post-corrective-statements-resolves-historic.

R. Uuk. Three Lines of Defence Against AI Risks at a Societal Level. Technical report, Future of Life Institute, 2023. URL https://ristouuk.com/wp-content/uploads/2023/06/Three_Lines_of_Defence_Against_AI_Risks_at_a_Societal_Level.pdf.

C. Vaccari and A. Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1):205630512090340, Jan. 2020. ISSN 2056-3051, 2056-3051. doi: 10.1177/2056305120903408. URL http://journals.sagepub.com/doi/10.1177/2056305120903408.

S. A. Vaghefi, D. Stammbach, V. Muccione, J. Bingler, J. Ni, M. Kraus, S. Allen, C. Colesanti-Senni, T. Wekhof, T. Schimanski, et al. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480, 2023.

S. Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, Oct. 2016. ISBN 9780190498511. URL https://global.oup.com/academic/product/technology-and-the-virtues-9780190498511?cc=gb&lang=en&.

G. Van de Kerk and A. R. Manuel. A comprehensive index for a sustainable society: The SSI—the sustainable society index. *Ecological Economics*, 66(2-3):228–242, 2008.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, Sept. 2016. URL http://arxiv.org/abs/1609.03499. arXiv:1609.03499 [cs].

W. van der Maden, D. Lomas, and P. Hekkert. Positive AI: Key challenges for designing wellbeing-aligned artificial intelligence, 2023.

A. van der Plas, M. Smits, and C. Wehrmann. Beyond Speculative Robot Ethics: A Vision Assessment Study on the Future of the Robotic Caretaker. *Accountability in Research*, 17(6):299–315, Nov. 2010. ISSN 0898-9621, 1545-5815. doi: 10.1080/08989621.2010.524078. URL https://www.tandfonline.com/doi/full/10.1080/08989621.2010.524078.

J. A. van Dijk. Digital divide research, achievements and shortcomings. *Poetics*, 34(4-5):221–235, Aug. 2006. ISSN 0304422X. doi: 10.1016/j.poetic.2006.05.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0304422X06000167.

T. J. VanderWeele. On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, 114(31):8148–8156, 2017.

V. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, Sept. 1999. ISSN 1941-0093. doi: 10.1109/72.788640. URL https://ieeexplore.ieee.org/abstract/document/788640.

J. Varelius. On the prospects of collective informed consent. *Journal of applied philosophy*, 25(1):35–44, 2008.

L. Z. Varga. Solutions to the routing problem: towards trustworthy autonomous vehicles. *Artificial Intelligence Review*, 55(7):5445–5484, Oct. 2022. ISSN 1573-7462. doi: 10.1007/s10462-021-10131-y. URL https://doi.org/10.1007/s10462-021-10131-y.

P. Vassilakopoulou and E. Hustad. Bridging Digital Divides: a Literature Review and Research Agenda for Information Systems Research. *Information Systems Frontiers*, 25(3):955–969, June 2023. ISSN 1572-9419. doi: 10.1007/s10796-020-10096-3. URL https://doi.org/10.1007/s10796-020-10096-3.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. Long Beach, CA, USA, 2017. arXiv. doi: 10.48550/arXiv.1706.03762. URL https://arxiv.org/pdf/1706.03762.pdf. arXiv:1706.03762 [cs].

G. Vella. Persuasion: importance of trust, relevance for small states, and limitations of computers - Diplo Resource, Aug. 2013. URL https://www.diplomacy.edu/resource/persuasion-importance-of-trust-relevance-for-small-states-and-limitations-of-computers/.

A. W. Vemuri and R. Costanza. The role of human, social, built, and natural capital in explaining life satisfaction at the country level: Toward a national well-being index (NWI). *Ecological economics*, 58(1):119–133, 2006.

P. Verma. ChatGPT provided better customer service than his staff. He fired them. *Washington Post*, Oct. 2023a. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/10/03/ai-customer-service-jobs/.

P. Verma. They fell in love with ai bots. a software update broke their hearts. *The Washington Post*, 2023b.

J. Vincent. Microsoft's Bing is an emotionally manipulative liar, and people love it, Feb. 2023a. URL https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams.

J. Vincent. Stack Overflow survey finds developers are ready to use AI tools — even if they don't fully trust them, June 2023b. URL https://www.theverge.com/2023/6/13/23759101/stack-overflow-developers-survey-ai-coding-tools-moderators-strike.

D. Vollrath. Will AI cause explosive economic growth?, July 2023. URL https://dietrichvollrath.substack.com/p/will-ai-cause-explosive-economic.

R. Von Schomberg. A Vision of Responsible Research and Innovation. In R. Owen, J. Bessant, and M. Heintz, editors, *Responsible Innovation*, pages 51–74. Wiley, 1 edition, Apr. 2013. ISBN 9781119966364 9781118551424. doi: 10.1002/9781118551424.ch3. URL https://onlinelibrary.wiley.com/doi/10.1002/9781118551424.ch3.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, Mar. 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/10.1126/science.aap9559.

V. Voukelatou, L. Gabrielli, I. Miliou, S. Cresci, R. Sharma, M. Tesconi, and L. Pappalardo. Measuring objective and subjective well-being: Dimensions and data sources. *International Journal of Data Science and Analytics*, 11:279–309, 2021.

C. Véliz. *Privacy is Power: Why and How You Should Take Back Control of Your Data*. Penguin Random House, July 2021. URL https://www.penguin.co.uk/books/442343/privacy-is-power-by-carissa-veliz/9780552177719.

C. Véliz. Chatbots shouldn't use emojis. *Nature*, 615(7952):375–375, Mar. 2023. doi: 10.1038/d41586-023-00758-y. URL https://www.nature.com/articles/d41586-023-00758-y.

S. Wachter and B. Mittelstadt. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, 2019:494, 2019. URL https://heinonline.org/HOL/Page?handle=hein.journals/colb2019&id=506&div=&collection=.

K. Wagner and H. Schramm-Klein. Alexa, Are You Human? Investigating Anthropomorphism of Digital Voice Assistants – A Qualitative Approach. *ICIS 2019 Proceedings*, Nov. 2019. URL https://aisel.aisnet.org/icis2019/human_computer_interact/human_computer_interact/7.

M. Wairagkar, M. R. De Lima, M. Harrison, P. Batey, S. Daniels, P. Barnaghi, D. J. Sharp, and R. Vaidyanathan. Conversational artificial intelligence and affective social robot for monitoring health and well-being of people with dementia. *Alzheimer's & Dementia*, 17: e053276, 2021.

J. Wajcman. *Pressed for time: The acceleration of life in digital capitalism*. University of Chicago Press, 2020.

L. Walker. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times*, Mar. 2023. URL https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal Adversarial Triggers for Attacking and Analyzing NLP, Jan. 2021. URL http://arxiv.org/abs/1908.07125. arXiv:1908.07125 [cs].

G. M. Walton and T. D. Wilson. Wise interventions: Psychological remedies for social and personal problems. *Psychological review*, 125(5): 617, 2018.

E. W. Wan and R. P. Chen. Anthropomorphism and object attachment. *Current Opinion in Psychology*, 39:88–93, June 2021. ISSN 2352250X. doi: 10.1016/j.copsyc.2020.08.009. URL https://linkinghub.elsevier.com/retrieve/pii/S2352250X20301548.

Q. Wan, S. Hu, Y. Zhang, P. Wang, B. Wen, and Z. Lu. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models, Aug. 2023. URL http://arxiv.org/abs/2307.10811. arXiv:2307.10811 [cs].

G. Wang, J. Zhao, M. Van Kleek, and N. Shadbolt. Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children. In *CHI Conference on Human Factors in Computing Systems*, pages 1–29, New Orleans LA USA, Apr. 2022a. ACM. ISBN 9781450391573. doi: 10.1145/3491102.3502057. URL https://dl.acm.org/doi/10.1145/3491102.3502057.

H. Wang. Transparency as Manipulation? Uncovering the Disciplinary Power of Algorithmic Transparency. *Philosophy & Technology*, 35(3): 69, July 2022. ISSN 2210-5441. doi: 10.1007/s13347-022-00564-w. URL https://doi.org/10.1007/s13347-022-00564-w.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, Nov. 2022b. URL http://arxiv.org/abs/2211.00593. arXiv:2211.00593 [cs].

R. E. Wang and D. Demszky. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction, June 2023. URL http://arxiv.org/abs/2306.03090. arXiv:2306.03090 [cs].

W. Wang. Smartphones as Social Actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior*, 68: 334–344, Mar. 2017. ISSN 07475632. doi: 10.1016/j.chb.2016.11.022. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563216307634.

X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good, Jan. 2020. URL http://arxiv.org/abs/1906.06725. arXiv:1906.06725 [cs].

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53 (3):1–34, May 2021. ISSN 0360-0300, 1557-7341. doi: 10.1145/3386252. URL https://dl.acm.org/doi/10.1145/3386252.

F. R. Ward, T. Everitt, F. Belardinelli, and F. Toni. Honesty is the best policy: defining and mitigating ai deception. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

C. Wardle and H. Derakhshan. Information Disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe, Sept. 2017. URL https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

N. Warford, T. Matthews, K. Yang, O. Akgul, S. Consolvo, P. G. Kelley, N. Malkin, M. L. Mazurek, M. Sleeper, and K. Thomas. Sok: A framework for unifying at-risk user research. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2344–2360, 2022. doi: 10.1109/SP46214.2022.9833643.

S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review*, 4(5):193, Dec. 1890. ISSN 0017811X. doi: 10.2307/1321160. URL https://www.jstor.org/stable/1321160?origin=crossref.

T. Warren. Microsoft has been secretly testing its Bing "Sydney" chatbot for years, Feb. 2023. URL https://www.theverge.com/2023/2/23/23609942/microsoft-bing-sydney-chatbot-history-ai.

M. Warschauer. Reconceptualizing the Digital Divide. *First Monday*, 7(7), July 2002. ISSN 13960466. doi: 10.5210/fm.v7i7.967. URL http://journals.uic.edu/ojs/index.php/fm/article/view/967.

M. Warschauer. *Technology and social inclusion: rethinking the digital divide*. MIT Press, Cambridge, Mass., 1. mit press paperback ed edition, 2004. ISBN 9780262731737.

M. Warschauer and T. Matuchniak. New Technology and Digital Worlds: Analyzing Evidence of Equity in Access, Use, and Outcomes. *Review of Research in Education*, 34(1):179–225, Mar. 2010. ISSN 0091-732X, 1935-1038. doi: 10.3102/0091732X09349791. URL http://journals.sagepub.com/doi/10.3102/0091732X09349791.

I. M. Wasserman and M. Richmond-Abbott. Gender and the Internet: Causes of Variation in Access, Level, and Scope of Use *. *Social Science Quarterly*, 86(1):252–270, Mar. 2005. ISSN 0038-4941, 1540-6237. doi: 10.1111/j.0038-4941.2005.00301.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.0038-4941.2005.00301.x.

A. Watts. New index measures well-being and ranks kentucky. *Foresight*, 11(1), 2004.

A. Waytz, J. Cacioppo, and N. Epley. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, May 2010. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691610369336. URL http://journals.sagepub.com/doi/10.1177/1745691610369336.

A. Waytz, J. T. Cacioppo, R. Hurlemann, F. Castelli, R. Adolphs, and L. K. Paul. Anthropomorphizing without Social Cues Requires the Basolateral Amygdala. *Journal of Cognitive Neuroscience*, 31(4):482–496, Apr. 2019. ISSN 0898-929X, 1530-8898. doi: 10.1162/jocn_a_01365. URL https://direct.mit.edu/jocn/article/31/4/482-496/28993.

N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold. ClimateBert: A pretrained language model for Climate-Related text. Oct. 2021.

A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How Does LLM Safety Training Fail?, July 2023a. URL http://arxiv.org/abs/2307.02483. arXiv:2307.02483 [cs].

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, Oct. 2022. doi: 10.48550/arXiv.2206.07682. URL http://arxiv.org/abs/2206.07682. arXiv:2206.07682 [cs].

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Jan. 2023b. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from Language Models, Dec. 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

L. Weidinger, M. G. Reinecke, and J. Haas. Artificial moral cognition: Learning from developmental psychology. preprint, PsyArXiv, Aug. 2022a. URL https://osf.io/tnf4e.

L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA, June 2022b. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://dl.acm.org/doi/10.1145/3531146.3533088.

L. Weidinger, K. R. McKee, R. Everett, S. Huang, T. O. Zhu, M. J. Chadwick, C. Summerfield, and I. Gabriel. Using the Veil of Ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18):e2213709120, May 2023a. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2213709120. URL https://pnas.org/doi/10.1073/pnas.2213709120.

L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023b.

J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang. Challenges in Detoxifying Language Models, Sept. 2021. URL http://arxiv.org/abs/2109.07445. arXiv:2109.07445 [cs].

C. Welch, C. Gu, J. K. Kummerfeld, V. Pérez-Rosas, and R. Mihalcea. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, 2022.

T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. A Network-based End-to-End Trainable Task-oriented Dialogue System. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1042.pdf.

Z. Wen, X. Geng, and Y. Ye. Does the use of WeChat lead to subjective well-being?: The effect of use intensity and motivations. *Cyberpsychology, Behavior, and Social Networking*, 19(10):587–592, 2016.

P. S. Wenz. *Environmental justice*. SUNY series in environmental public policy. State University of New York Press, Albany, 1988. ISBN 9780887066443 9780887066450.

P. H. Werhane. *Moral imagination and management decision-making*. Oxford University Press, USA, 1999.

P. H. Werhane. Moral imagination and systems thinking. *Journal of business ethics*, 38:33–42, 2002.

A. Wertheimer. *Exploitation*. Princeton University Press, Aug. 1999. ISBN 9780691019475. URL https://press.princeton.edu/books/paperback/9780691019475/exploitation.

M. West, R. Kraut, and H. E. Chew. I'd blush if I could: closing gender divides in digital skills through education. Technical report, UNESCO, Jan. 2019. URL https://unesdoc.unesco.org/ark:/48223/pf0000367416.

J. Whittaker, S. Looney, A. Reed, and F. Votta. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10 (2), June 2021. ISSN 2197-6775. doi: 10.14763/2021.2.1565. URL https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content.

M. Whittaker, M. Alper, C. L. Bennett, S. Hendren, L. Kaziunas, M. Mills, M. Ringel Morris, J. Rankin, E. Rogers, M. Salas, and S. Myers West. Disability, Bias, and AI. Technical report, AI Now Institute, Nov. 2019. URL https://ainowinstitute.org/publication/disabilitybiasai-2019.

J. Whittlestone and J. Clark. Why and how governments should monitor ai development, 2021.

P. H. Wicksteed. *The common sense of political economy, including a study of the human basis of economic law*. Macmillan, 1910.

K. Widner, S. Virmani, J. Krause, J. Nayar, R. Tiwari, E. R. Pedersen, D. Jeji, N. Hammel, Y. Matias, G. S. Corrado, Y. Liu, L. Peng, and D. R. Webster. Lessons learned from translating AI from development to deployment in healthcare. *Nature Medicine*, 29(6):1304–1306, June 2023. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-023-02293-9. URL https://www.nature.com/articles/s41591-023-02293-9.

J. B. Wiener. The regulation of technology, and the technology of regulation. *Technology in Society*, 26(2-3):483–500, Apr. 2004. ISSN 0160791X. doi: 10.1016/j.techsoc.2004.01.033. URL https://linkinghub.elsevier.com/retrieve/pii/S0160791X04000375.

K. Wiggers and A. Stringer. ChatGPT: Everything you need to know about the AI chatbot, Nov. 2023. URL https://techcrunch.com/2023/11/06/chatgpt-everything-to-know-about-the-ai-chatbot/.

J. W. Wiktor and K. Sanak-Kosmowska. *Information asymmetry in online advertising*. Routledge studies in marketing. Routledge, Milton Park, Abingdon, Oxon ; New York, NY, 2021. ISBN 9781000454055 9781003134121. URL https://www.routledge.com/Information-Asymmetry-in-Online-Advertising/Wiktor-Sanak-Kosmowska/p/book/9780367652128.

B. Williams. *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press, Princeton, Dec. 2010. ISBN 9781400825141. doi: 10.1515/9781400825141. URL https://www.degruyter.com/document/doi/10.1515/9781400825141/html.

J. Williams, A. Raux, D. Ramachandran, and A. Black. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France, Aug. 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-4065.pdf.

T. Williamson. *Vagueness*. Routledge, New York, 1994.

T. Williamson. Précis of Vagueness. *Philosophy and Phenomenological Research*, 57(4):921–928, 1997. ISSN 0031-8205. doi: 10.2307/2953810. URL https://www.jstor.org/stable/2953810.

S. Willison. Bing: "I will not harm you unless you harm me first", Feb. 2023. URL https://simonwillison.net/2023/Feb/15/bing/.

H. H. Wilmer, L. E. Sherman, and J. M. Chein. Smartphones and Cognition: A Review of Research Exploring the Links between Mobile Technology Habits and Cognitive Functioning. *Frontiers in Psychology*, 8:605, Apr. 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00605. URL http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00605/full.

C. Wilson. Is it love or loneliness? Exploring the impact of everyday digital technology use on the wellbeing of older adults. *Ageing & Society*, 38(7):1307–1331, 2018.

T. Wilson. *Redirect: The surprising new science of psychological change*. Penguin UK, 2011.

L. Winner. *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press, 2010.

C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

J. C. Witte and S. E. Mannon. *The Internet and social inequalities*. Contemporary sociological perspectives. Routledge, New York, 2010. ISBN 9780415963206 9780415963190 9780203861639. OCLC: ocn166361305.

W. Wong, P. Dutta, O. Voicu, Y. Chervonyi, C. Paduraru, and J. Luo. Optimizing Industrial HVAC Systems with Hierarchical Reinforcement Learning, Sept. 2022. URL http://arxiv.org/abs/2209.08112. arXiv:2209.08112 [cs, eess].

A. W. Wood. Coercion, Manipulation, Exploitation. In C. Coons and M. Weber, editors, *Manipulation*, pages 17–50. Oxford University Press, Aug. 2014. ISBN 9780199338207. doi: 10.1093/acprof:oso/9780199338207.003.0002. URL https://academic.oup.com/book/4870/chapter/147239348.

S. C. Woolley. Automating power: Social bot interference in global politics. *First Monday*, Mar. 2016. ISSN 1396-0466. doi: 10.5210/fm.v21i4.6161. URL https://journals.uic.edu/ojs/index.php/fm/article/view/6161.

B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. De Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. Von Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li,

D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. McDuff, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. De Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, June 2023. URL http://arxiv.org/abs/2211.05100. arXiv:2211.05100 [cs].

World Health Organization. COP24 special report: health and climate change. Technical report, World Health Organization, 2018. URL https://www.who.int/publications-detail-redirect/9789241514972.

World Inequality Database. World Inequality Database. URL https://wid.world/.

C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. A. Behram, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. Lee, H.-H. S. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. Hazelwood. Sustainable AI: Environmental Implications, Challenges and Opportunities, Jan. 2022. URL http://arxiv.org/abs/2111.00364. arXiv:2111.00364 [cs].

S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. NExT-GPT: Any-to-Any Multimodal LLM, Sept. 2023. URL http://arxiv.org/abs/2309.05519. arXiv:2309.05519 [cs].

T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021.acl-long.523.

Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan. See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–9, Oldenburg Germany, Oct. 2020. ACM. ISBN 9781450375160. doi: 10.1145/3379503.3403563. URL https://dl.acm.org/doi/10.1145/3379503.3403563.

C. D. L. Wynne. The perils of anthropomorphism. *Nature*, 428(6983):606–606, Apr. 2004. ISSN 0028-0836, 1476-4687. doi: 10.1038/428606a. URL https://www.nature.com/articles/428606a.

C. Xiang. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says, Mar. 2023. URL https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

T. Xie and I. Pentina. Attachment theory as a framework to understand relationships with social chatbots: a case study of replika. 2022.

A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein. Detoxifying Language Models Risks Marginalizing Minority Voices, Apr. 2021a. URL http://arxiv.org/abs/2104.06390. arXiv:2104.06390 [cs].

J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan. Bot-Adversarial Dialogue for Safe Conversational Agents. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online, June 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL https://aclanthology.org/2021.naacl-main.235.

J. Xu, A. Szlam, and J. Weston. Beyond Goldfish Memory: Long-Term Open-Domain Conversation, July 2021c. URL https://arxiv.org/pdf/2107.07567.pdf. arXiv:2107.07567 [cs].

P. Xu, X. Zhu, and D. A. Clifton. Multimodal Learning with Transformers: A Survey, May 2023. URL http://arxiv.org/abs/2206.06488. arXiv:2206.06488 [cs].

M. M. Yamin, M. Ullah, H. Ullah, and B. Katt. Weaponized AI for cyber attacks. *Journal of Information Security and Applications*, 57: 102722, Mar. 2021. ISSN 2214-2126. doi: 10.1016/j.jisa.2020.102722. URL https://www.sciencedirect.com/science/article/pii/S2214212620308620.

R. Yang, X. Sun, and K. Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in Neural Information Processing Systems*, 32, 2019.

J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. Enhanced Membership Inference Attacks against Machine Learning Models, Sept. 2022. URL http://arxiv.org/abs/2111.09679. arXiv:2111.09679 [cs, stat].

D. T. Yeong Tan and R. Singh. Attitudes and Attraction: A Developmental Study of the Similarity-Attraction and Dissimilarity-Repulsion Hypotheses. *Personality and Social Psychology Bulletin*, 21(9):975–986, Sept. 1995. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167295219011. URL http://journals.sagepub.com/doi/10.1177/0146167295219011.

A. Ymous, K. Spiel, O. Keyes, R. M. Williams, J. Good, E. Hornecker, and C. L. Bennett. "I am just terrified of my future" Epistemic Violence in Disability Related Technology Research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Honolulu HI USA, Apr. 2020. ACM. ISBN 9781450368193. doi: 10.1145/3334480.3381828. URL https://dl.acm.org/doi/10.1145/3334480.3381828.

K. Yogeeswaran, J. Złotowski, M. Livingstone, C. Bartneck, H. Sumioka, and H. Ishiguro. The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, 5(2):29–47, 2016.

X. Yue, H. A. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, and R. Sim. Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe, July 2023. URL http://arxiv.org/abs/2210.14348. arXiv:2210.14348 [cs].

T. Zahavy, B. O'Donoghue, A. Barreto, V. Mnih, S. Flennerhag, and S. Singh. Discovering diverse nearly optimal policies with successor features. *arXiv preprint arXiv:2106.00669*, 2021.

T. Zahavy, Y. Schroecker, F. Behbahani, K. Baumli, S. Flennerhag, S. Hou, and S. Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. *arXiv preprint arXiv:2205.13521*, 2022.

T. Zahavy, V. Veeriah, S. Hou, K. Waugh, M. Lai, E. Leurent, N. Tomasev, L. Schut, D. Hassabis, and S. Singh. Diversifying AI: Towards creative chess with alphazero, 2023.

M. Zajko. Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3):e12962, Mar. 2022. ISSN 1751-9020, 1751-9020. doi: 10.1111/soc4.12962. URL https://compass.onlinelibrary.wiley.com/doi/10.1111/soc4.12962.

M. Zalnieriute and T. Cutts. How AI and New Technologies Reinforce Systemic Racism. Technical report, United Nations Human Rights Council, Oct. 2022. URL https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/advisorycommittee/study-advancement-racial-justice/2022-10-26/HRC-Adv-comm-Racial-Justice-zalnieriute-cutts.pdf.

P. Zeinert, N. Inie, and L. Derczynski. Annotating Online Misogyny. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.247. URL https://aclanthology.org/2021.acl-long.247v2.pdf.

B. Zevenbergen. *Internet users as vulnerable and at-risk human subjects: reviewing research ethics Law for technical internet research*. http://purl.org/dc/dcmitype/Text, University of Oxford, 2020. URL https://ora.ox.ac.uk/objects/uuid:5363f3f8-6c13-4e56-b065-2980bdcce46b.

J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen. Recommendation as instruction following: A large language model empowered recommendation approach, 2023.

Q. Zhang, J. Lu, and Y. Jin. Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1):439–457, Feb. 2021. ISSN 2198-6053. doi: 10.1007/s40747-020-00212-w. URL https://doi.org/10.1007/s40747-020-00212-w.

Q. Zhu and J. Luo. Generative Design Ideation: A Natural Language Generation Approach, Mar. 2022. URL http://arxiv.org/abs/2204.09658. arXiv:2204.09658 [cs].

E. Zhuravskaya, M. Petrova, and R. Enikolopov. Political effects of the internet and social media. *Annual review of economics*, 12:415–438, 2020.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-Tuning Language Models from Human Preferences, Jan. 2020. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs, stat].

A. Zimmermann, K. Vredenburgh, and S. Lazar. The Political Philosophy of Data and AI. *Canadian Journal of Philosophy*, 52(1):1–5, Jan. 2022. ISSN 0045-5091, 1911-0820. doi: 10.1017/can.2022.28. URL https://www.cambridge.org/core/product/identifier/S0045509122000285/type/journal_article.

S. Zuboff. *The Age of Surveillance Capitalism*. Hachette, June 2017. ISBN 9781610395694. URL https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/?lens=publicaffairs.

M. Zwolinski, B. Ferguson, and A. Wertheimer. Exploitation. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition, 2022. URL https://plato.stanford.edu/archives/win2022/entries/exploitation/.