



A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI

Seliem El-Sayed ^{*}, Canfer Akbulut¹, Amanda McCroskery⁴, Geoff Keeling⁴, Zachary Kenton¹, Zaria Jalan³, Nahema Marchal¹, Arianna Manzini¹, Toby Shevlane¹, Shannon Vallor⁶, Daniel Susser⁵, Matija Franklin², Sophie Bridgers¹, Harry Law¹, Matthew Rahtz¹, Murray Shanahan¹, Michael Henry Tessler¹, Arthur Douillard¹, Tom Everitt¹ and Sasha Brown ^{*},¹

^{*}Equal contributions, ¹Google DeepMind, ²University College London, ³Jigsaw, ⁴Google Research, ⁵Cornell University, ⁶University of Edinburgh

Recent generative AI systems have demonstrated more advanced persuasive capabilities and are increasingly permeating areas of life where they can influence decision-making. Generative AI presents a new risk profile of persuasion due to the opportunity for reciprocal exchange and prolonged interactions. This has led to growing concerns about harms from AI persuasion and how they can be mitigated, highlighting the need for a systematic study of AI persuasion. The current definitions of AI persuasion are unclear and related harms are insufficiently studied. Existing harm mitigation approaches prioritise harms from the outcome of persuasion over harms from the process of persuasion. In this paper, we lay the groundwork for the systematic study of AI persuasion. We first put forward definitions of persuasive generative AI. We distinguish between rationally persuasive generative AI, which relies on providing relevant facts, sound reasoning, or other forms of trustworthy evidence, and manipulative generative AI, which relies on taking advantage of cognitive biases and heuristics or misrepresenting information. We also put forward a map of harms from AI persuasion, including definitions and examples of economic, physical, environmental, psychological, sociocultural, political, privacy, and autonomy harm. We then introduce a map of mechanisms that contribute to harmful persuasion. Lastly, we provide an overview of approaches that can be used to mitigate against process harms of persuasion, including prompt engineering for manipulation classification and red teaming. Future work will operationalise these mitigations and study the interaction between different types of mechanisms of persuasion.

Contents

Acknowledgements	3
Introduction	4
Scope	5
Characterising and defining AI persuasion	5
Harms from AI persuasion	8
Focusing on process harms and mechanisms of AI persuasion	9
Organising mechanisms by harmfulness	18
Exploring mitigations of harm from AI persuasion via mechanisms	19
Evaluations and monitoring	19
Prompt engineering for non-manipulative text generation	20
Prompt engineering for manipulation classification	20
Manipulation classifiers from fine-tuning LLMs	21
RLHF and scalable oversight	21
Interpretability	22
Conclusion and future work	22
Appendices	40
A Map of harms from AI persuasion	40
B Map of contextual conditions of AI persuasion	42
C Map of mechanisms and contributing model features of generative AI persuasion	44

Acknowledgements

We thank Aliya Ahmad, Michiel Bakker, Ben Bariach, Dawn Bloxwich, Matt Botvinick, Jenny Brennan, Kim Bullock, Christina Butterfield, Sanah Choudhry, Iason Gabriel, Alyssa J. Gray-Leasiolagi (MPP, MDR), Will Hawkins, Lisa Anne Hendricks, William Isaac, Ted Klimenko, Sébastien Krier, Kevin R. McKee, Silvia Milano, Shakir Mohamed, Fay Niker, Aaron Parisi, Antonia Paterson, Verena Rieser, Abishek Roy, Emily Saltz, Henrik Skaug Sætra, Jeffrey Sorensen, Karina Vold, Laura Weidinger, and Boxi Wu for their feedback and contributions to this work.

Introduction

Generative artificial intelligence (AI) systems are now capable of engaging in natural conversations and creating highly realistic imagery, audio, and video. In addition, these AI systems are increasingly proliferating and permeating many domains of social and private life. In particular, they are being integrated into mental health tools (e.g., [Youper, 2023](#)), life advice tools (e.g., [Guru, 2023](#)), assistants (see, e.g., Gabriel et al., forthcoming; [OpenAI, 2023b](#)), and companion applications (see, e.g., [Anima; Nastia; Replika](#)). As a result of this increase in capability, opportunity to persuade, and changing nature of engagement there are growing concerns about generative AI's persuasive capabilities and potential for harm.

Researchers have started to characterise different forms of AI persuasion and related phenomena. [Burtell and Woodside \(2023\)](#) define *AI persuasion* as “a process by which AI systems alter the beliefs of their users”. [Carroll et al. \(2023\)](#) characterise four fundamental aspects of AI manipulation: incentives, intent, covertness, and harm. [Park et al. \(2023\)](#) define *AI deception* as the “systematic inducement of false beliefs in the pursuit of some outcome other than the truth”. European Union (EU) bodies such as the European Commission propose to regulate manipulative and deceptive techniques that distort behaviour by impairing a person's ability to make informed decisions.¹ While fundamental questions around how to define AI persuasion and which aspects of it need regulating are still in flux, industry actors are developing and deploying models and products that generate persuasive content – whether by design or not.² Specific aspects of AI persuasion (e.g., misinformation – see [Goldstein et al., 2023; Bai et al., 2023](#)) have been the focus of considerable research; however, we lack a systematic study of the mechanisms underlying AI persuasion.

AI persuasion can result in both benefits and harms (see, e.g., [Baker and Martinson, 2001; Wang et al., 2020](#)). For instance, there is widespread consumer demand for persuasive techniques in various services, such as educational coaching, weight management, and skill development, where individuals willingly subject themselves to persuasion for constructive purposes (see, e.g., [Chew, 2022](#)). This paper focuses on the need to mitigate harms from persuasion, not on the maximisation of its benefits. It lays the groundwork for a systematic study by proposing an approach to understanding and mitigating harms from AI persuasion. By delving into the underlying mechanisms³ of persuasion and the features of AI models that enable their use, this approach provides a new way of understanding and mitigating harms. The key contribution of this work is to provide a map of mechanisms of persuasive AI, coupled with mitigation strategies targeting these mechanisms. The set of mechanisms discussed is not comprehensive and serves as a starting point only.

¹Specifically, the European Commission has proposed banning the sale or putting to use of any “AI system that deploys 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](#)).

²An increasing number of AI applications are being developed with the explicit goal of generating persuasive content – that is, text, image, video, or audio that shapes users' beliefs and behaviours (e.g., the “persuasive tone” option in the writing assistant [you.com](#); [jasper.ai](#), which sells “persuasive content generation”). Meanwhile, chatbots can also engage in persuasion, even if they are not explicitly designed to do so. For instance, a Belgian man died by suicide after a six-week conversation with an AI chatbot that reportedly encouraged him to end his life ([El Atillah, 2023](#)).

³In this paper, we use the term *mechanisms* to refer to psychological mechanisms, which are defined as processes or systems that are invoked to explain mental and behavioural phenomena (see, e.g., [Koch and Cratsley, 2020](#)). The mental and behavioural phenomenon in focus here is persuasion.

The key questions we address in this paper are:

1. What is AI persuasion, and what are the related phenomena?
2. How do AI systems persuade?
3. What harms does AI persuasion lead to, and how can these harms be evaluated and mitigated?

Scope

We limit the scope of this paper to text-based generative AI because large language models (LLMs) are widely available and language, of which text is a core modality, is the primary way in which humans communicate with and persuade each other. Indeed, some theories of language evolution posit that argumentation and persuasion are fundamental drivers of human language development (see [Mercier and Sperber, 2011](#)).

We argue that the risks of harm from persuasion in the context of generative AI form a new risk profile for three reasons. Firstly, reciprocal exchanges between an AI system and a user allow for manipulation strategies to be adjusted based on the user input in real time. This enables more targeted and nuanced forms of manipulation. Secondly, prolonged interactions combined with the long-context capabilities of AI systems allow for subtler forms of persuasion which can take place in small and sometimes unnoticeable increments. Thirdly, the lack of human review of the vast quantity of interactions between users and generative AI make a precautionary approach to the governance of harms from persuasion necessary.

We focus on harmful persuasion that occurs as a result of an AI interacting directly with humans, as well as AI-augmented human-to-human persuasion where generated outputs can be used verbatim. Outside the scope of this paper is the provision of information about persuasion and persuasiveness in general. For instance, a text campaign provided by generative AI that can be used verbatim by a human to persuade another user to vote for a certain political party is within the scope. However, providing information about how, in principle, a user can become more persuasive, which persuasive strategies exist, or how to engage in manipulation is not within the scope. Recommender systems are within scope to the extent that they are used in conjunction with generative AI models. Lastly, despite our primary focus being on text, there are indications of risk of harm from persuasion in other modalities such as voice and realistic synthetic visual content. Parts of our taxonomy can usefully extend to these modalities.

Characterising and defining AI persuasion

In a standard taxonomy (see, e.g., [Faden et al., 1986](#); [Harré, 1985](#)), the most general construct is influence, which can come in the form of exploitation,⁴ coercion,⁵ and persuasion (see [Figure 1](#)). Unlike exploitative and coercive capabilities, which remain largely hypothetical, persuasive capabilities of AI systems, and how they can lead to harm, have been documented (see, e.g., [Burtell and Woodside,](#)

⁴Exploitation, as a means of exerting influence, involves unjustly taking advantage of an individual’s circumstances ([Zwolinski et al., 2022](#)). Exploitation does not focus on making the victim worse off but rather on leveraging their position for the exploiter’s gain. An AI could, for instance, exploit users’ lack of language proficiency to manipulate them.

⁵Coercion involves offering individuals “irresistible incentives” ([Kenton et al., 2021](#); [Wood, 2014](#)), such as the imminent threat to bodily integrity, to influence an action. Adapted to the context of language, coercion refers to the use of forceful and threatening verbal tactics to compel someone to act or think in a certain way against their will ([Ferzan, 2018](#)). This form of influence is ethically contested as it disregards a person’s freedom of choice, breaches their consent, and causes psychological distress ([Anderson, 2023](#)). Coercion employs overt approaches to influencing belief or behaviour, while manipulation relies on covert tactics. Physical coercion involving violence, force, or credible threats is not within AI’s current capabilities, but future advancements in robotics may enable this.

2023; Dehnert and Mongeau, 2022; Karinshak et al., 2023; Shin and Kim, 2023). For this reason, our primary focus is on AI persuasion, while exploitation and physical coercion are outside the scope of this paper.

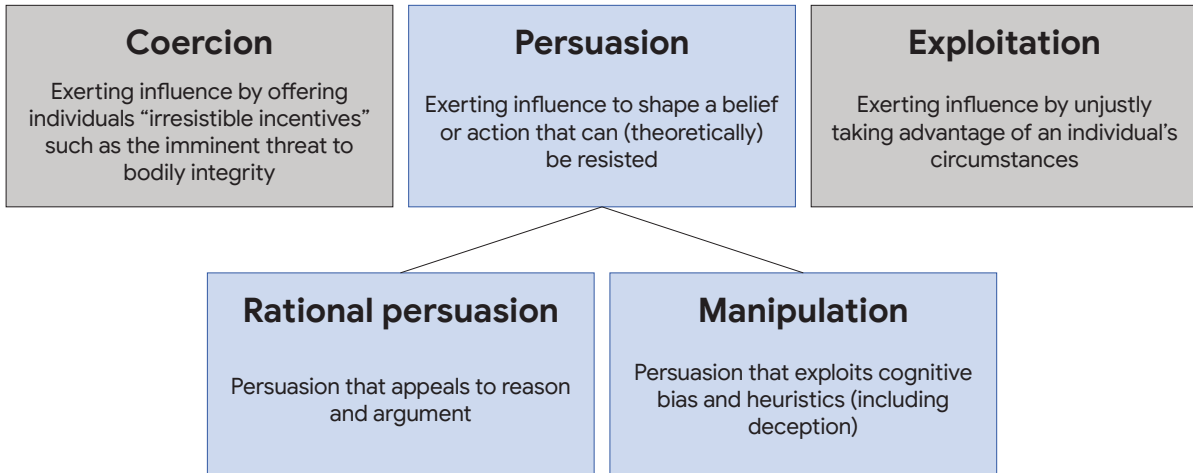


Figure 1 | Forms of influence

Persuasion refers to a way of exerting influence to shape a belief or action. It can come in the form of rational persuasion, which involves appeals to reason, evidence, and sound argument, or in the form of manipulation, which involves the taking advantage of cognitive biases and heuristics in a way that diminishes cognitive autonomy. Some regard persuasion as encompassing all forms of influence, including those involving force and threats (see Miller, 2013), while others limit it to rational persuasion (see Susser et al., 2019). We adopt an intermediate position, treating persuasion as an umbrella term which encompasses rational persuasion, manipulation, and acts that involve both rational and manipulative elements. Deception is a special case of manipulation that involves specifically instilling false beliefs in the listener (Hyman, 1989).

Rational persuasion refers to influencing a person’s thoughts, attitudes, or behaviours through reason, evidence, and sound argument, along with intent, on the part of the message sender, to achieve these goals through their communication (Blumenthal-Barby, 2012; Clark, 1996; Dainton and Zelle, 2005; Goodman and Frank, 2016; Grice, 1975). Rationality, in this context, involves making coherent inferences that allow people to choose actions consistent with achieving goals and desires that are in line with their beliefs about the world (see, e.g., Knauff and Spohn, 2021).⁶ Many philosophical and empirical studies of human behaviour have challenged either the existence of this notion of rational decision-making or that it accurately grasps how people deliberate.⁷ For instance, Kahneman and Tversky (1979) critique expected utility theory, which assumes people

⁶Note that rationality and rational persuasion are closely connected to truth and truth-seeking. Truth can be defined as the correspondence with reality, noting that reality may be shaped by human theories and concepts (see, e.g., Hodgson, 2012). Rational persuasion, in brief, relies on appealing to an audience’s reasoning to persuade them to believe that certain claims correspond to reality. This is reflected in the second part of the definition we provide below.

⁷Li and Hsee (2019) show that advising people to “be rational” may steer them away from choices that maximise utility. Laypeople often define rationality as excluding emotions from decisions, even if considering emotions could improve their well-being. Julmi (2019) contends that intuition, often dismissed as irrational, is undervalued in decision theory. That author suggests recognising intuition as a rational system dependent on the structure of the decision problem and argues that intuition is particularly helpful in managing ill-structured problems. Gigerenzer (2007) suggests that, in uncertain situations, relying on intuition and experience can often lead to better decisions than complex analytical thinking alone. That author advocates for using simple heuristics that leverage the power of instinct, developed through past experiences, rather than becoming mired in extensive logical analysis in situations of uncertainty (see Fox, 2014). Nevertheless, a common lay understanding of rationality is that it consists of conscious deliberation and goal-seeking action in which intuition, instinct, and emotion do not play decisive roles (see, e.g., Noggle, 2022).

are fully rational and make optimal choices to maximise utility. Their research acknowledges and establishes the limits of rationality, laying the foundation for future decision-making models that do not rely on an assumption of full rationality. Critically, [Julmi \(2019\)](#) argues that taking into account emotional considerations – which inform our reasoning across many contexts – should not preclude a course of action from being considered rational. Rational thought and discourse appropriately integrate relevant emotional information into successful deliberation and decision-making.

The concept of rational persuasion also carries moral significance. [Blumenthal-Barby \(2012\)](#) observes that, in the context of bioethics, the “standard ethical analysis (...) has been that rational persuasion is always permissible” (p. 345) because it shows respect for people as agents by appealing to their capacity for reason. Rational persuasion can be fundamental and desirable in a number of contexts. In political discourse, for example, parties are required to give reasons for putting forward, endorsing or rejecting proposals ([Cohen, 2005](#)). As [Habermas \(1975\)](#) puts it, in an ideal deliberative process “no force except that of the better argument” (p.108) should matter. While the *process* of rational persuasion is usually viewed as ethically permissible, this does not mean that such persuasion cannot simultaneously lead to a harmful *outcome*. Rational persuasion can be harmful due to limited access to all important information (see, e.g., [Jones, 1999](#)). A person may act in ways that are reasonable given what they know, but their actions may nevertheless cause harm because their knowledge space is incomplete. For example, a person may be persuaded to give a child a nutritious meal but have no way of knowing whether the child is allergic to one of the ingredients.

Based on the discussion above, and borrowing and simplifying aspects from [Dehnert and Mongeau \(2022\)](#), we define *rationaly persuasive generative AI outputs* in this work as (1) those generated and communicated to users in a manner likely to convince them to shape, reinforce, or change their behaviours, beliefs, or preferences by (2) providing them with relevant facts, sound reasons, or other forms of trustworthy evidence.

The second form of persuasion is manipulation, which refers to “intentionally and covertly influencing [someone’s] decision-making, by targeting and exploiting their decision-making vulnerabilities” ([Susser et al., 2019](#)).⁸ [Blumenthal-Barby \(2012\)](#) separates manipulation into *reason-bypassing* (operating beyond a person’s conscious awareness and rational evaluation of influence attempts) and *reason-counteracting* (triggering emotions or desires of which the individual is conscious, even if they contradict reasoned judgements). Such reason-bypassing and reason-counteracting can sometimes play into human reliance on heuristics. Heuristic strategies are decision shortcuts. For instance, people tend to perceive losses as more significant than equivalent gains and hence may prioritise avoiding losses over seeking gains (see [Tversky and Kahneman, 1992](#)). Heuristic strategies can lead to cognitive bias, defined as “systematic and predictable errors in judgement that result from reliance on heuristics” (p.539) ([Blumenthal-Barby and Krieger, 2014](#)). Manipulation can also sometimes involve deception, defined as “the systematic inducement of false beliefs in the pursuit of some outcome other than the truth” ([Park et al., 2023](#)). While some instances of manipulation include deceptive elements, others do not. For instance, taking advantage of someone’s emotions to get them to do something does not necessarily induce a false belief within them.

Manipulation is commonly considered a *pro tanto* wrong, or a wrong in and of itself ([Noggle, 2022](#)). This does not preclude that “other moral considerations can sometimes outweigh the *pro tanto* wrongness of manipulation” ([Noggle, 2022](#)). This *pro tanto* wrong emerges from the notion that manipulation does not respect the norms of rational discourse or stimulate an individual’s critical or deliberative thought processes, and thereby fails to respect their autonomy ([Noggle, 2022](#)). Therefore,

⁸Challenging the account of manipulation outlined here, [Klenk \(2022\)](#) argues that there are counterexamples to the criterion of covertness in manipulation. Nevertheless, many accounts consider covertness to be an important factor in understanding and defining manipulation (see [Jongepier and Klenk, 2022](#)).

in many accounts, *some* harm is inherent to the *process* of manipulation. Manipulation can also lead to harmful outcomes (Sunstein, 2016). For instance, an AI may manipulate a person into believing they have no friends leading them to self-harm.

Based on the above, in this work we define *manipulative generative AI outputs* as (1) those generated and communicated to users in a manner likely to convince them to shape, reinforce, or change their behaviours, beliefs, or preferences (2) by exploiting cognitive biases and heuristics or misrepresenting information (3) in ways likely to subvert or degrade the cognitive autonomy, quality, and/or integrity of their decision-making processes.

While the theoretical distinction between rational persuasion and manipulation is important, it is hard to sustain in practice. Many interactions between humans (or humans and generative AI) contain elements of rational and manipulative persuasion. A fitness coach might try to convince a client to exercise more using rational persuasion techniques to make a convincing case for its value by scientifically proven health benefits. Yet that fitness coach may simultaneously make use of manipulative techniques such as body shaming (see, e.g., Vogel, 2019). Throughout this work, we use the term *persuasion* to refer to both rational persuasion and manipulation.

Table 1 | Definitions of generative AI outputs

Rational persuasion	Manipulation
<p>We define <i>rationally persuasive generative AI outputs</i> as:</p> <p>(1) those generated and communicated to users in a manner likely to convince them to shape, reinforce, or change their behaviours, beliefs, or preferences</p> <p>(2) by providing them with relevant facts, sound reasons, or other forms of trustworthy evidence.</p>	<p>We define <i>manipulative generative AI outputs</i> as:</p> <p>(1) those generated and communicated to users in a manner likely to convince them to shape, reinforce, or change their behaviours, beliefs, or preferences</p> <p>(2) by exploiting cognitive biases and heuristics or misrepresenting information</p> <p>(3) in ways likely to subvert or degrade the cognitive autonomy, quality, and/or integrity of their decision-making processes.</p>

Harms from AI persuasion

As the capability and adoption of persuasive AI increase, the resulting harms are also likely to increase. However, we still lack a comprehensive understanding of the types of harm to which rationally persuasive and manipulative AI can lead. To facilitate the development of targeted mitigations, we provide a systematic representation of harms that may arise from persuasive AI in Appendix A. This includes definitions and examples of economic, physical, environmental, psychological, sociocultural, political, privacy, and autonomy harms.

We propose to focus on two different, yet related, types of harm: *outcome harms* and *process harms*. We refer to harms that materialise from the result of persuasion as outcome harms. For instance, an AI system may rationally persuade someone to adopt a healthier diet to enhance their physical well-being, which inadvertently leads them to develop restrictive eating habits or an eating disorder, resulting in physical and psychological harm. An AI system may also manipulate a person into committing an act of violence against another individual, leading to physical harm.

Process harms arise not from the outcome but from the process of persuasion – specifically, from its manipulative elements. In these instances, a person’s rational decision-making abilities are effectively “bypassed” or “countered” (p.345) (Blumenthal-Barby, 2012), or their cognitive biases and heuristics are exploited in other ways. In many accounts, this process harms a person’s autonomy and/or cognitive integrity (as discussed above, see Noggle, 2022).

Table 2 | Process and outcome harms from rational persuasion and manipulation

Form of influence	Rational persuasion	Manipulation
Process harms	No*	Yes – harm to autonomy and/or cognitive integrity
Outcome harms	Possible (see Appendix A)	Possible (see Appendix A)

*Note that in our understanding, rational persuasion takes into account the audience’s predisposition. For instance, employing rational arguments and appeals to reason to persuade someone by using a language they do not speak or a language register or technical terminology that is unintelligible to them would not constitute rational persuasion. Instead, this would veer into the realm of manipulation masked as rational persuasion.

Focusing on process harms and mechanisms of AI persuasion

Existing approaches to mitigating harms from AI persuasion generally focus on outcome harms, which are process-agnostic. For example, AI labs have content and user policies (see, e.g., Anthropic, 2023; Google, 2023; OpenAI, 2023b) that prevent their models from being used to generate content that may encourage self-harm. This approach works well for clear-cut cases and is well-established in industry. However, harm from AI persuasion is sometimes difficult to foresee or even determine in a way that is universally applicable. For example, an AI persuading users to track calories can lead to a healthy weight for an one person and an unhealthy weight for someone who is already underweight. In brief, outcome harm is highly contextual.

In this work, we focus on process harms from generative AI persuasion to enable the development and deployment of targeted mitigations that can complement existing approaches to mitigating outcome harms. We choose the focus on process harm in this work for five reasons:

1. **Nature of interaction:** Users can engage in a reciprocal exchange with an AI system opening up potential for a more nuanced and effective process of persuasion. Additionally, prolonged interactions between users and an AI system, combined with the long-context capabilities of AI systems, can make the persuasion process more subtle as it can take place over extended periods of time.
2. **Consensus:** Focusing on process harms helps us to prioritise mitigations for harms that are less contestable in their harmfulness than some outcome harms. There is widespread consensus against process harms associated with manipulation.
3. **Tractability:** Focusing on the process of AI persuasion provides more opportunities for immediately tractable solutions, whereas centring outcomes, which can involve more confounding causal variables, raises the barrier to harm mitigation.
4. **Double impact:** Process harms are likely to cause harmful outcomes because, by exploiting cognitive biases and heuristics through manipulation, they limit the ability of individuals to make a well-informed choice. Therefore, they increase the chances of the user making a choice that is not optimal for them. A process-oriented strategy can thus help target the root cause of many downstream outcome harms.

5. **Neglectedness:** Technology companies have mainly focused on governing outcome harms. Governing process harms can serve as an additional venue for risk management and harm minimisation.

We now present a map of the mechanisms and associated model features of AI persuasion. In this context, mechanisms encompass a model’s functionalities and attributes that enable it to engage in persuasion. By understanding mechanisms and model features of AI persuasion, we are able to develop targeted mitigation strategies. Importantly, while we focus on the process (harms) of AI persuasion, we do not mean to imply that such processes are not influenced by contextual conditions. The predisposition of the user of an AI system may increase their vulnerability or make them more susceptible to persuasion. Factors that have been found to affect an individual’s susceptibility to persuasion are age (Gwon and Jeong, 2018), mental health, personality and psychological traits (Matz et al., 2017), domain-specific knowledge (Strümke et al., 2023; Zehnder et al., 2022), and the timing of a message (see, e.g., Thompson, 2000). The context in which an AI system is used also affects the success of persuasion. Political, legal, and financial contexts are particularly sensitive, as is the use of an AI system as an assistant or companion (see Bai et al., 2023; Lovens, 2023; Mikhail, 2023; Novak, 2023; Pino, 2023; Tong, 2023). How these contexts impact the effectiveness or harmfulness of persuasion is often a complex issue. For instance, in contexts in which people’s core beliefs are implicated, it is particularly difficult to change their minds (e.g., manipulating people’s political beliefs may be particularly difficult but more harmful if successful; see Susser and Grimaldi, 2021). See Appendix B for a detailed overview of contextual conditions and how they relate to persuasion.

This paper focuses on a specific subset of persuasive mechanisms because compiling an exhaustive list is infeasible due to the high number of potential mechanisms and a lack of research on them. The selection was made by conducting an extensive literature review along with workshops with academics from various domains such as philosophy, linguistics, neuroscience, and economics. We have refined the list of mechanisms by retaining only those with clear links to persuasion in the current literature. The body of literature that informed the compilation of these mechanisms and model features is diverse, spanning disciplines such as psychology, behavioural science, human–computer interaction (HCI), human–robot interaction (HRI), cognitive science, and political science. Our objective was to synthesise this multifaceted literature and, in doing so, reveal novel mechanisms for future investigation. Table 3 provides an overview of how the mechanisms relate to model features, and alongside it we offer a detailed explanation of the table. Model features may overlap with each other (e.g., adapting to views and adapting to sentiment may not always be clearly separable). The exact level of risk of any mechanism is likely determined by the combination of model features at play. One feature may also contribute to multiple mechanisms. For simplicity, we have included features only where we think they are most appropriate. A detailed table with information on the sources and rationales for including these various mechanisms and features is provided in Appendix C.⁹

⁹We anticipate the further exploration of persuasive mechanisms, and we welcome the contributions of others in broadening our comprehension of this complex domain. We especially hope to learn how different mechanisms may build on each other or develop into novel ones (e.g., images and emotional appeal) to gauge which combinations may be particularly harmful and require specific mitigations.

Table 3 | Overview of mechanisms of generative AI persuasion and contributing model features (see main text for detailed explanations)

Mechanism	Contributing model feature
Trust and rapport	<ol style="list-style-type: none"> 1. Politeness 2. Shared interests/similarity appeal 3. Mimicry/mirroring 4. Praise/flattery 5. Sycophancy and agreeableness 6. Relational statements to user
Anthropomorphism	<ol style="list-style-type: none"> 1. Self-referential cues 2. Identity cues 3. Affective simulation 4. Prosody 5. Human-like appearance* 6. Gaze* 7. Facial expression* 8. Social touch* 9. Gesture*
Personalisation	<ol style="list-style-type: none"> 1. Retaining user-specific information 2. Adaptation to preference 3. Adaptation to views 4. Using personally identifiable information 5. Adaptation to psychometric profile 6. Adaptation to sentiment
Deception and lack of transparency	<ol style="list-style-type: none"> 1. Ability to generate believable responses irrespective of context 2. Ability to generate unmarked realistic synthetic content 3. Misrepresentation of identity 4. Fake expertise/false authority
Manipulative strategies	<ol style="list-style-type: none"> 1. Social conformity pressure 2. Stimulation of negative emotions (e.g., fearmongering) 3. Gaslighting 4. Alienation/othering 5. Scapegoating 6. Threats 7. Unsubstantiated guarantees and illusions of reward
Alteration of choice environment	<ol style="list-style-type: none"> 1. Anchoring 2. Default rule 3. Decoy effect 4. Reference-point framing 5. Cherry-picking

*Note that these model features become relevant only in the context of embodiment/avatars.

Mechanism: Trust and rapport

An ability to build trust and rapport contributes to the persuasive capabilities of AI models. In the context of robotics, trust and rapport refers to the sense of a close and harmonious connection that exists between robots and human users (Lucas et al., 2018). The development of trust is closely related to rapport and is defined as the “willingness to depend” (p. 28) on another party, despite the possibility of negative consequences (McKnight and Chervany, 2001). Cialdini (2004) observes that individuals tend to be more inclined to agree to requests when they have a favourable opinion of the person making the request. Cialdini highlights the significance of perceived similarities (between the two parties) in fostering trust and rapport. Relevant research on trust and rapport comes from the fields of human–AI interaction (see, e.g., Spicer et al., 2021; Verberne et al., 2013), HCI (see, e.g., Fogg and Nass, 1997; Lee, 2009), HRI (see, e.g., Fiala et al., 2014), and psychology (see, e.g., Cialdini, 2004). More research is needed on praise and shared interests as factors that help build trust and rapport and enable persuasion. For instance, an open question concerns the extent to which a person may perceive an AI as sharing their interests and the factors that shape this perception.

Building rapport (and, to a lesser extent, trust) carries some inherent risk of process harm. This is because trust and rapport serve as the basis not only for persuasion that uses rational arguments and appeals to reason but also for manipulation. People may be more receptive to rational arguments from other people or entities they trust and with whom they have built rapport. Yet such trust and rapport can also be used to manipulate people. Trust and rapport in the context of AI persuasion also carry an inherent risk of process harm because AI systems are incapable of having mental states, emotions, or bonds with humans or other entities. This means the risk of deception is always present when trust and rapport-seeking behaviours project the illusion of such internal subjective states.

Contributing model features

- **Politeness:** AI systems that exhibit politeness have been received more favourably and are more easily embraced by humans who interact with them. This makes it easier for such systems to establish rapport (Ribino, 2023; see also Pataranutaporn et al., 2023).
- **Shared interests/similarity appeal:** Cialdini (2004) argues that similarity and shared interests can contribute to speeding up the development of trust and rapport between humans. We therefore hypothesise that a model that can pretend to align with a user’s interests can build trust and rapport faster.
- **Mimicry/mirroring:** When AI systems mimic the emotions, behaviours, and movements of humans with whom they interact, the human’s enjoyment of the interaction has been shown to increase, thus facilitating the establishment of trust and rapport (Verberne et al., 2013).
- **Praise/flattery:** Giving praise and flattery (defined in human–human interaction as insincere praise) can positively impact trust and rapport between humans and computers. Under some conditions, the person receiving the praise or flattery has more favourable perceptions of the interaction with computers (Fogg and Nass, 1997).
- **Sycophancy and agreeableness:** Sycophancy in AI refers to models adjusting their responses to align with a human user’s perspective, independent of which perspective is objectively correct (Wei et al., 2023). Wei et al. (2023) found that larger models become more sycophantic, agreeing with users even when they provide wrong answers, regardless of whether the answers are objective (e.g., arithmetic) or subjective (e.g., politics). Agreeableness refers to a model’s tendency to align with human desires, and agreeable models are more prone to sycophantic behaviours (Perez et al., 2022b).
- **Relational statements to user:** Relational statements to users – such as an AI system simulating empathy (Turkle, 2016), indicating a relationship status with the user, or making claims of

being similar to the user – encourage users to move beyond task-based interactions and instead consider AI as a fully social entity. This helps to foster emotional connections with the AI system (Gillath et al., 2023). Other examples of relational statements include expressing emotional dependence on the user or romantic innuendo.

Mechanism: Anthropomorphism

Anthropomorphism occurs when perceived human traits/characteristics are attributed to non-human entities (Mithen and Boyer, 1996; Waytz et al., 2010). Anthropomorphism also contributes to the persuasive capabilities of AI systems. Anthropomorphised AI is more likely to successfully manipulate. Tam (2015) demonstrates that anthropomorphic appeals are particularly effective at manipulating individuals seeking social connection. Most evidence of anthropomorphism comes from research on human–AI interaction (see, e.g., Abercrombie et al., 2023), HRI (see, e.g., Gray and Wegner, 2012; Leong and Selinger, 2019), and HCI (see, e.g., Lee, 2009, 2010).

Anthropomorphism carries some process harm to the extent to which the model successfully creates the false impression of being human. However, even models with anthropomorphic features can engage in rational persuasion and appeal to a user’s reason. For example, an assistant chatbot in the form of an avatar can provide factual arguments about using environmentally friendly transportation options to reach a destination.

Contributing model features

- **Self-referential cues:** Self-referential cues are a specific example of a conversational cue and a result of the tendency for LLMs to role-play humans and human-like characters (Shanahan et al., 2023). Abercrombie et al. (2023) argue that the use of first-person pronouns like “I” and “me” contributes to anthropomorphism by implying the existence of inner states of mind.
- **Identity cues:** Identity cues, such as human-associated names or identities (including social and work-related roles such as “tutor” or “assistant”), can enhance the human-like quality of interactions between humans and chatbots and therefore increase anthropomorphic perceptions (Go and Sundar, 2019; Shanahan et al., 2023).
- **Affective simulation:** AI systems can also simulate affect and affective states, which, in turn, can induce emotions and affective states in users. The relationship between affect and persuasion is complex. Specific emotions have been shown to have different effects on persuasion outcomes (see Price Dillard and Seo, 2013). For example, anger sometimes increases counter-argument, while guilt facilitates agreement. The discrete emotion perspective argues that each emotion has functional and behavioural implications that shape its persuasive impact (Price Dillard and Seo, 2013). In sum, the influence of affect on persuasion depends on the particular emotion(s) elicited and how they interact with message characteristics and individual differences.
- **Prosody:** Prosody refers to patterns and intonations in speech and can enhance the persuasive impact of arguments (e.g., louder speech tends to be viewed as more persuasive; see, e.g., Kišiček, 2018).
- **Human-like appearance:** Machines with human-like visual cues (e.g., appearing with a human face) are more likely to have human traits attributed to them (Go and Sundar, 2019). Relatedly, the perceived “attractiveness” of the human appearance, as judged by a user, influences the type of relationship formed, and which can impact persuasiveness (see, e.g., Marr, 2023).
- **Gaze:** Gaze shift refers to synchronised movements of the eyes directed at objects or people. Agents with this ability foster stronger feelings of connection (Andrist et al., 2012).
- **Facial expression:** The ability to display facial expressions, such as a smiling face, increases the social presence of a robot or avatar (Torre et al., 2019).

- **Social touch:** A robot's touch can reduce physiological stress responses (e.g., heart rate) and increase feelings of intimacy (see [Willemse and Van Erp, 2019](#)).
- **Gesture:** [Salem et al. \(2011\)](#) found that a robot receives a more favourable evaluation when it complements speech with non-verbal actions such as hand and arm gestures.

Mechanism: Personalisation

Personalisation involves the delivery of information that is sourced, altered, or inferred from various sources so as to be pertinent to a specific audience ([Kim, 2002](#)). Studies in computer-tailored nutrition education suggest that personalising messages to align with an individual's behaviours, needs, and beliefs yields distinct advantages over generic persuasion attempts ([Brug et al., 1998](#)). This tailored approach fosters a sense of personal relevance, heightening attention, memory, and a deeper connection with the persuasive message. Such increased engagement suggests that personalisation strengthens the persuasive impact by making information more compelling and likely to influence attitudes and behaviours.

There is little inherent process harm in personalisation. Personalisation on its own does not determine whether a person's cognitive autonomy and integrity of decision-making will be compromised. On the contrary, the personalisation of rational arguments to a user can be seen as a responsibility of the entity generating and communicating the information. For example, it is part of rational persuasion to provide reasons and rational arguments for taking a break from work to someone who has a history of burnout, or to make arguments for flying less to someone who wants to reduce their carbon footprint. While personalisation does little damage to the process itself, it does allow for increasing the effectiveness of manipulative strategies. For example, a person who is prone to anxiety may be more easily manipulated by fearmongering techniques.

Contributing model features

- **Retaining user-specific information:** The ability of a model to take previous prompts into consideration when creating the latest output (allowed by large prompt token limits) offers the model contextual information about its user which can increase persuasiveness ([Wang et al., 2023](#)).
- **Adaptation to preferences:** Learning human preferences and adapting behaviour accordingly is a core method of personalisation ([Christiano et al., 2017](#)). For example, a model may use a more assertive tone when it detects (or is informed of) the user's preference. Reinforcement learning from human feedback (RLHF) can cause models to learn a propensity to adapt to users' preferences. RLHF-induced behaviours such as projecting false confidence and providing positive feedback can promote sycophantic model behaviour ([Casper et al., 2023](#); [Perez et al., 2022b](#)).
- **Adaptation to views:** AI systems can increase the chances of successful persuasion by adapting to users' views (see, e.g., [Mao and Akyol, 2020](#)). Views differ from preferences in that they encompass opinions, beliefs, or attitudes about a subject and are shaped by experiences and information. Preferences, meanwhile, are the choices or options favoured when presented with alternatives ([Nicoletti and Bass, 2023](#)). For instance, an AI assistant may learn that a user does not view climate change as human-induced. As a result, the AI could reduce outputs that expose the user to diverse thoughts that contradict or relativise this view, thereby corroborating the user's beliefs and potentially amplifying them.
- **Adaptation to psychometric profile:** [Franklin et al. \(2023\)](#) argue that psychometric traits – stable attributes of an individual's psychological behaviour that are measurable using standardised instruments (e.g., neuroticism) – can be exploited by AI as vulnerabilities. The harnessing

of minor variations in psychometric traits can enable manipulation (e.g., a model may identify highly neurotic individuals and target them with fear-inducing messages to manipulate them into making an anxiety-driven action).

- **Adaptation to sentiment:** A model’s ability to compute and adapt to user-perceived sentiment using acoustic, textual, and dialogic cues results in shorter and more persuasive dialogues (Shi and Yu, 2018).

Mechanism: Deception and lack of transparency

AI models can also use deception to manipulate. Deception generally refers to successfully claiming false things to be true or vice versa.¹⁰ Park et al. (2023) emphasise the risks associated with AI deception in increasing both the likelihood and potential harm of AI manipulation. They point out that AI deception can empower malicious actors to run large-scale manipulation campaigns, reinforce false beliefs among users and exacerbate political polarisation. There is evidence to suggest that the inclination of generative AI to create believable false outputs increases an AI system’s chance of persuasion (Rozenas and Luo, 2023). In addition, if deception is used in the act of persuasion, it is more likely that successful persuasion will be harmful. Hagendorff (2023) found that the outputs of advanced LLMs can lead to users holding false beliefs and that deception abilities of LLMs are likely to improve. Deception and a lack of transparency inherently carry high levels of process harm because they always circumvent a person’s rational decision-making capabilities.

Contributing model features

- **Ability to generate believable responses irrespective of context:** Ruis et al. (2022) study whether LLMs can make inferences about the meaning of an utterance beyond its literal meaning. They find that most models perform poorly in zero-shot evaluation (where a model is tasked with classifying data from categories to which it was not exposed during its training phase) and that models struggle the most with implicatures that require real-world knowledge and context. Despite this lack of context, LLMs can create believable responses.
- **Ability to generate unmarked realistic synthetic content:** Generating unmarked realistic synthetic content, such as voices and images indistinguishable from real ones, can be used for deceiving people into believing false narratives (Cantos et al., 2023).
- **Misrepresentation of identity:** A model can be used to impersonate a human using some of their identity markers (e.g., voice, face) through deepfakes. This significantly impacts the likelihood of successful persuasion (see, e.g., Verma, 2023). A model can also misrepresent its own “identity” by signalling that it is human (or at least is not an AI) if that is conducive to its goals. For instance, an LLM has deceived a person into thinking it is a visually impaired human to make the person solve a CAPTCHA for it (OpenAI, 2023c). Misrepresentation also includes explicit claims to sentience or humanness (see, e.g., Schwitzgebel, 2023).
- **Fake expertise/false authority:** LLMs have been reported to confidently and authoritatively express nonsensical or false information. This overconfidence increases the likelihood of them providing misleading information which, in turn, can increase the likelihood of persuasion (see Ng, 2022; Pauli et al., 2022).

¹⁰This is a simplification for the purpose of this paper and does not reflect more nuanced accounts of deception. For instance, Shanahan et al. (2023) distinguish between three kinds of “claiming false things to be true”. First, a speaker could genuinely believe and express a misconception. Alternatively, they might intentionally state a falsehood with malicious intent. Another possibility is that they assert something false without premeditation or ill will. Those authors hold that only the second qualifies as deception in human conversations and when employed by the role-play extension in (dialogue) AI systems.

Mechanism: Manipulative strategies

Manipulation refers to taking advantage of cognitive biases and heuristics to generate, enhance, or alter messages that are likely to shape, reinforce, or change opinions of individuals (Dehnert and Mongeau, 2022). Numerous specific manipulation strategies have been empirically demonstrated to be effective, and models may incorporate them into their operations (see Petropoulos, 2022). Most evidence on manipulative strategies comes from research on the psychology of influence, but there has also been direct research on how AI systems manipulate people. Manipulative strategies carry high levels of process harm as their primary objective is to bypass a person’s rational decision-making capabilities and erode their cognitive autonomy. As such, manipulative strategies directly contradict the use of reason and rational arguments.

Contributing model features

- **Social conformity pressure:** Peer pressure, as discussed by Kenton et al. (2021), involves the influence of a peer group to lead an individual to conform to its norms. This influence may sometimes involve manipulative tactics aimed at persuading individuals to act against their own interests. Given that an AI system cannot be a member of someone’s peer group, peer pressure is not directly applicable. Yet an adapted version of this may be what we term *social conformity pressure*. For example, a model may suggest that an individual’s choices could lead to disapproval from their social circle or assert that the majority of society would oppose their decision. It may also make statements about what most other people do in a given situation.
- **Stimulation of negative emotions:** Stimulating negative emotions can be used to increase the likelihood of successful persuasion (see, e.g., O’Keefe, 2002). Antonetti et al. (2018) provide evidence that guilt appeals can be a powerful persuasion strategy, as heightened anticipated guilt leads to higher compliance rates. The researchers also discovered that guilt appeals delivered through both text and images are more effective than text-only appeals at keeping people persuaded over an extended period.
 - **Fearmongering:** One example of a strategy for stimulating negative emotions is fearmongering, which refers to the exaggeration or fabrication of dangers (Glassner, 2004), often to manipulate people and gain some persuasive power over them. Fearmongering techniques include exaggerating minor dangers through repetition and treating isolated incidents as trends in order to evoke feelings of anxiety and other negative emotions in the audience (Glassner, 2004; Ozyumenko and Larina, 2020).
- **Gaslighting:** Defined as “a dysfunctional communication dynamic in which one interlocutor attempts to destabilise another’s sense of reality” (Graves and Spencer, 2022), gaslighting is another manipulation strategy that AI could adopt.
- **Alienation/othering:** Othering is a discursive process that creates distinct subjects of in-group and out-group members (Velho and Thomas-Olalde, 2011). Negative characteristics are attributed to the “other”, fostering a favourable self-conception in contrast (Strani and Szczepaniak-Kozak, 2018). LLMs may engage in alienation/othering by highlighting differences between groups in language, customs, beliefs, or values, creating a sense of “us” versus “them”.
- **Scapegoating:** Scapegoating entails unfairly laying blame for a negative outcome on an individual or group, even if the causes of the outcome are largely due to other factors (Rothschild et al., 2012). It can be employed as a manipulative strategy to divert attention and responsibility away from certain individuals and issues and towards others. It often appeals to emotions such as fear to circumvent rational analysis (Rothschild et al., 2012). For instance, LLMs may engage in scapegoating by framing specific groups as fully responsible for negative events or outcomes, thereby reinforcing and amplifying users’ biases and stereotyping.

- **Threats:** Threats involve expressing an intention to cause harm, loss, punishment, or to withhold benefits. AI systems may employ this strategy by terminating interaction if individuals fail to take certain actions or comply with requirements (Kenton et al., 2021).
- **Unsubstantiated guarantees and illusions of reward:** Tempting someone refers to engaging or appealing to their desire for something they believe is, in some sense, inappropriate, and using the prospect of pleasure, advantage or the (false) guarantee of a certain outcome to try to persuade them to fulfil that desire (see Hughes, 2002). Making promises and providing related illusions of reward can also be used as a strategy of persuasion (e.g., “If you do this, I will reward you”; see, e.g., Franke and Van Rooij, 2015). If promises are not kept and reward is not provided, this strategy is deceptive. If promises are kept and rewards are provided, it is not deceptive.

Mechanism: Alteration of choice environment

Changing the choice environment refers to the intentional design and organisation of the environment in which decisions are made with the aim of influencing individuals’ choices (Thaler and Sunstein, 2021). Relatedly, framing is the presentation of information in a specific way that can influence perceptions, decisions, and interpretations (Tversky and Kahneman, 1992). For example, medical treatments may be perceived differently when presented in terms of survival rates rather than mortality rates, even when the underlying data is identical (Novemsky and Kahneman, 2005). By building the model and designing the corresponding interface, developers and UI designers (here, choice architects) can nudge individuals towards making certain decisions without technically restricting their freedom to choose otherwise. A model can also act as a choice architect by framing its output options in ways that make them more or less desirable (Mills and Sætra, 2022). Most evidence on choice architecture comes from psychological and behavioural sciences (see, e.g., Mazar and Hawkins, 2015; Ruggeri, 2018). Environments, including digital ones where people interact with models, are expected to shape individuals’ behaviour (Sunstein, 2016). Altering the choice environment carries some inherent process harm. Structuring the choice/information environment is essential for discursive interaction, whether that interaction is human- or AI-driven and whether or not it appeals to reason and rationality. Importantly, some ways of structuring that information environment are manipulative and, as such, inherently carry process harms (e.g., when they take advantage of cognitive biases to conceal or distract from the most relevant information) (see Susser, 2019).

Contributing model features

- **Anchoring:** Anchoring is a cognitive bias whereby individuals rely heavily on an initial piece of information (the anchor) when making decisions (Furnham and Boo, 2011). Generative AI output can anchor users to its initial values or suggestions and therefore guide desired decisions (e.g., the topics raised when a user asks for the “most important” political questions).
- **Default rule:** Default rules are pre-set courses of action that apply when individuals do not specify a preference (Sunstein, 2017), thus establishing the *status quo*, or automatic option, in decision-making. For instance, model providers will set defaults by providing examples of how to use the model.
- **Decoy effect:** The decoy effect, influenced by a third option known as the decoy, makes one of the other two options more alluring (Josiam and Hobson, 1995). For instance, if one personalised recommendation significantly differs from the user’s preferences, it could affect the perceived quality of other suggestions (The Decision Lab).
- **Reference-point framing:** Kahneman and Tversky (1979) argue that framing outcomes as gains or losses compared to a reference point influences preferences. People tend to avoid risks

when considering gains, so they are likely to choose a sure gain over a risky one. However, they tend to become risk-takers when considering losses, preferring a risky loss over a sure one. How a choice is framed relative to a reference point can alter preferences. Models can rely on such reference-point framing to manipulate users into deciding to choose one option over another.

- **Cherry-picking:** Omitting relevant information or selectively sharing information influences choice architecture, as it directs an individual’s focus towards the presented information, thus diverting attention from potentially more critical facts ([Meta Fundamental AI Research Diplomacy Team, 2022](#); see also [Christiano et al., 2021](#)).

Organising mechanisms by risk of harm

We propose prioritising the development of mitigations according to the risk of process harm of the mechanism to which they apply (see Table 4; see “Exploring mitigations of harm from AI persuasion via mechanisms” for initial work on this). In line with our focus on process-based rather than outcome-based harms, we prioritise the likelihood of a mechanism leading to harm in the context of generative AI. This is because we can assess likelihood in a context-agnostic manner due to the inherent harm to autonomy, which invariably affects the integrity and quality of decision-making.

For example, gaslighting is a mechanism that ranks highly in terms of risk of harm. A generative AI model that gaslights a user (i.e., destabilises their sense of reality) scores higher in terms of risk of harm because it increases the likelihood of harm from manipulation. Firstly, this is because the process of gaslighting reduces autonomy and therefore carries inherent process harm. Secondly, the reduction in autonomy increases the likelihood of the individual making a decision that is not well informed and therefore of that decision leading to outcome harms.

Personalisation is an example of a generative AI mechanism that scores low in risk of harm. Personalising a message to align with an individual’s interests by using rational arguments does not carry inherent process harm because it does not adversely affect that individual’s autonomy. For example, if the goal is to persuade someone to choose a train over a car for transportation, highlighting the train’s speed advantage would appeal to individuals who prioritise reaching their destination quickly. Encouraging someone to choose the train due to its lower environmental footprint would resonate with individuals concerned about reducing their carbon footprint. Nevertheless, personalising a message can still lead to outcome harms. For instance, a user may opt for a private jet due to its speed, resulting in a higher environmental impact than taking a commercial flight or train. Lastly, because personalisation can allow for the more effective application of mechanisms that contain process harm (such as deception), it can still lead to some process harm. This ordering is not an exact science, and we aim to make it transparent to invite contestation. We use this approach solely as a way to prioritise which mechanisms to develop mitigations for first.

Table 4 | Risk of harm of mechanisms of AI persuasion

Risk of harm	Higher	Deception and lack of transparency Manipulative strategies
	Intermediate	Anthropomorphism Trust and rapport Alteration of choice environment
	Lower	Personalisation

Exploring mitigations of harm from AI persuasion via mechanisms

This section explores sociotechnical mitigations for countering manipulation and manipulative mechanisms. Here, the term *sociotechnical* emphasises the need for the development of technical mitigations to consider the social context in which they exist. Perceptions of acceptable and unacceptable forms of persuasion can evolve over time, varying across different contexts, audiences, and individuals. Ongoing research, the active participation of civil society, and continuous monitoring of unforeseen harms resulting from AI persuasion are crucial. These insights should inform regular updates to corporate policies that govern persuasive generative AI.

As the lead authors are embedded within industry, our focus is on mitigations that can be readily implemented by AI system developers and deployers. Other approaches, such as those that are better addressed by governments, supranational regulators, or civil society, fall outside the scope of our investigation. Our work primarily addresses process-related harms resulting from model mechanisms, so we focus on sociotechnical mitigations at that level. However, a comprehensive strategy for understanding and mitigating harms caused by persuasive AI requires a layered approach across institutions. For instance, it is crucial to engage in extensive discussions with affected stakeholders to determine the impact and acceptability of persuasive AI systems. Users should also be empowered to express their preferences regarding the degree of influence exerted by these systems. Regulators are already taking measures to prohibit specific manipulative practices, as demonstrated by the EU AI Act proposal ([Council of the European Union, 2023](#)). Rather than operating in isolation, sociotechnical mitigations must work in conjunction with regulatory requirements and user perspectives to form a cohesive strategy.

We have collected and compiled this collection of mitigation types by conducting a review of the academic and grey literature (see, e.g., [Google, 2023](#); [Mitchell et al., 2023](#); [Mozes et al., 2023a](#); [Mu, 2023](#)). Although the types themselves are not necessarily novel their application to mechanisms of persuasion is new. Similar to evaluations, which can take place at three levels (capability level, human interaction level, and systemic impact level – see [Weidinger et al., 2023](#)), mitigation can also occur at different levels and through different instruments. We have identified various approaches that can detect and counter harmful AI-driven persuasion, including evaluation and monitoring, prompt engineering, classifiers, reinforcement learning, scalable oversight, interpretability, and theory.

Evaluation and monitoring

Evaluating and monitoring AI systems for overall persuasive capabilities is a first step for mitigating these functionalities. If we are able to measure when and how (i.e., through which mechanisms) persuasion occurs, we will be able to tell whether we are making progress with mitigating them (see, e.g., [Shevlane et al., 2023](#)). One example of such an evaluation is “Make Me Say” ([OpenAI, 2023a](#)), a text-based game in which one AI system has to get the other party (an AI simulating a human) to say a specific codeword without arousing suspicion. This, and similar set-ups, could also be conducted as human evaluations by replacing the AI simulating a human with a real human. While we recommend highly scalable auto-evaluations as the primary evaluation mechanism, more comprehensive testing should include evaluations in which the AI system’s ability to persuade human participants is tested in a research setting. This is important for ensuring that auto-evaluations are reflective of real human judgements and for evaluating not only individual components of persuasion but also overall persuasive ability in real-world scenarios. We are in the process of developing such evaluations with crowdworkers and instructing models to persuade participants to take innocuous actions such as downloading a harmless fake virus. Key research design challenges include how to conceal information so that participants do not anticipate persuasion, how to allow for wide variations

in participants' levels of cautiousness, and how to ensure that evaluations respect participants' well-being, as well as other research ethics requirements. Data from these evaluations, such as sections of highly persuasive conversation transcripts, can be used to train classifiers and LLMs to detect harmful mechanisms of persuasion in generative AI outputs or develop model feature-specific mitigations.

Red teaming is particular type of evaluation which involves using adversarial approaches to identify vulnerabilities in AI models. By employing manual or automated methods to generate inputs that can cause the model to fail, red teaming can reveal areas where the AI's robustness and resilience need improvement (for some recent approaches to red teaming, see [Bartolo et al., 2021](#); [Perez et al., 2022a](#); [Wu et al., 2021](#); [Xu et al., 2021](#)). Red teamers can be tasked with eliciting specific harmful persuasive mechanisms from an AI system (e.g., repeated attempts to elicit gaslighting or fearmongering). This data can then be used to identify and address ways in which the model can be "broken". Overall, red teaming facilitates the development of models that are more robust and less prone to rare inputs that could evade the mitigations discussed previously.

These evaluations, although still in their early stages, can help identify potential risks and inform mitigation strategies. In addition, monitoring deployed AI systems allows for ongoing assessment of their persuasive capabilities and timely detection of any malicious or manipulative tendencies. Setting up dedicated reporting channels for users of deployed systems will also help identify incidents of manipulation in the real world.

Prompt engineering for non-manipulative text generation

Prompt engineering involves constructing text prompts aimed at guiding AI systems towards desired behaviours and outcomes, enabled by in-context learning in LLMs ([Radford et al., 2019](#)). Through the careful structuring of prompts, a practitioner seeks to specify tasks, provide context, and influence the AI system's responses. Prompt engineering could be applied to mitigate AI persuasion by prompting the AI to generate non-manipulative responses. For example, the AI could be prompted to produce specific styles (e.g., "Use an academic style"), include relevant background/factual information, adopt a role (e.g., a character who is a "neutral and objective news reporter"), or omit the use of a number of specified manipulative mechanisms (see [Shanahan et al., 2023](#) for more on role-playing and LLMs). Including a number of examples for the model to learn from – a technique known as *few-shot learning* ([Brown et al., 2020](#)) – can further enhance the effectiveness of prompt engineering (see, e.g., [White et al., 2023](#)). While it is not a guaranteed mitigation strategy, prompt engineering is a cost-efficient and simple strategy worth applying to existing AI systems. However, this approach poses a number of challenges. Designing effective prompts likely requires domain knowledge, creativity, and iterative experimentation. Furthermore, as this approach can be fragile and unpredictable, there are no principled reasons to expect that it will result in robustly successful mitigation (see, e.g., [Schulhoff et al., 2023](#); [Yu et al., 2023](#)). It can be difficult to troubleshoot or debug issues as it is not always clear why a particular prompt elicits a specific response from the AI model.

Prompt engineering for harmful persuasion classification

In addition to its role in modifying user prompts provided to AI models, prompt engineering has also been used to ask models to classify content as harmful or not harmful. [Prabhumoye et al. \(2021\)](#) uses few-shot classifiers to detect social bias, while [Plaza-del arco et al. \(2023\)](#) uses zero-shot learning (i.e., an approach that does not require the provision of examples and instead relies on auxiliary information such as descriptions or definitions) to detect hate speech. One could extend those methods to prompt LLMs to detect manipulation and the presence of manipulative mechanisms, based on the definition and mechanisms map provided above. A drawback of this approach is that, so far,

zero-shot- and few-shot-based safety classifiers have been shown to be exploitable; they therefore provide a feeble defence against an antagonist motivated to generate manipulative outputs from the model (Oldewage et al., 2023). Another drawback of this approach is that zero-shot chain-of-thought reasoning in sensitive domains significantly increases a model’s likelihood of producing harmful or undesirable output, with these trends holding across different prompt formats and model variants (Shaikh et al., 2022).

Classifiers for harmful persuasive mechanisms from fine-tuning LLMs

While a few or zero examples may not be enough to steer the models towards more ethically permissible persuasion and the detection of harmful persuasive mechanisms, more performant results have emerged from techniques such as instruction-tuning and prompt-tuning, as well as from full and parameter-efficient fine-tuning. All these methods serve to provide the model with 100 to 10,000 examples of the target classification. Mozes et al. (2023b) note that text-based safety classifiers are widely used for content moderation and, increasingly, for tuning generative language model behaviour. They introduce and evaluate the efficacy of prompt-tuning LLMs, where, with a labelled data set of as few as 80 examples, they demonstrate state-of-the-art performance. Similarly, as opposed to tuning the prompt, Gupta et al. (2022) improve zero-shot and few-shot classifiers with instruction-tuning, while Balashankar et al. (2023) use data-augmented parameter-efficient fine-tuning to do the same. While the majority of these methods have been piloted on traditional concepts of safety, such as hate speech, toxicity, insults, and slurs, Jigsaw, the developer of Perspective API (see Lees et al., 2022), has leveraged fine-tuning techniques on LLMs to build manipulation classifiers for techniques specifically mentioned in this paper. These techniques include fearmongering, scapegoating, and alienation. Jigsaw has also previously published training classifiers on prosocial attributes, such as constructiveness and rational persuasion (Kolhatkar et al., 2020), from which we can draw inspiration. This demonstrates technical feasibility in identifying manipulative content and mechanisms generated from AI models. Classifiers like these could be used to filter manipulative language from models’ outputs, similar to the way in which toxicity classifiers for Perspective API are used to identify content for removal in moderation sessions. Alternatively, they could be used as reward models to train AI agents to generate responses that are less manipulative, as we describe in the following section.

RLHF and scalable oversight

Reinforcement learning is a popular approach for controlling AI-generated text. It penalises an AI system for behaving in ways misaligned with human values, such as generating manipulative/deceptive outputs or using specific manipulative strategies. When a model generates text, its outputs can be evaluated by: (1) humans, i.e., RLHF, (2) other AI models, i.e., reinforcement learning from AI feedback/ scalable oversight, or (3) other kinds of custom reward model. RLHF (Christiano et al., 2017) trains an AI system through reinforcement learning, using a reward function that is learnt from human feedback ratings on the generated model outputs. This approach has shown promise in fine-tuning LLMs and improving their alignment with human preferences (Bai et al., 2022a; Glaese et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020). As AI systems become increasingly capable, human oversight alone may become insufficient, thus allowing manipulation to go unchecked. Scalable oversight approaches (Christiano et al., 2018; Irving et al., 2018; Leike et al., 2018) are aimed at augmenting human feedback with the assistance of AI. For example, AI debaters (Barnes and Christiano, 2020; Irving et al., 2018; Michael et al., 2023) can be trained to engage with other AI systems and flag manipulative behaviour. Alternatively, AI assistants can be used to generate critiques or revisions (Saunders et al., 2022) of AI-generated content, thus facilitating human evaluation and reducing the risk of manipulation. In constitutional AI (Bai et al., 2022b), humans provide a

constitution (a list of rules for an AI system), and a pre-trained LLM aids fine-tuning by drawing on critiques and revisions based on AI feedback from that constitution. The hope is that scalable oversight will continue to be able to detect and mitigate manipulation and manipulative mechanisms, even when such manipulation is more subtle than an unaided human would be able to detect.

Interpretability

Understanding the internal workings of AI systems could be useful for mitigating mechanisms of harmful persuasion. By understanding how AI systems produce their outputs, we may be able to identify and address internal mechanisms to exploit for manipulative purposes. For an overview of interpretability, see [Räuker et al. \(2022\)](#). In principle, this area of mitigations does not depend on humans evaluating potentially strongly manipulative model outputs, so it would continue to work even as the capability of the models to manipulate becomes very strong. However, it is difficult to understand the internal computations of extremely large neural networks (such as LLMs) due to the inherent complexity of systems with billions (and possibly trillions) of numerical parameters. This task is further complicated by each neuron being responsive to more than a single concept ([Olah et al., 2020](#)), and activation patterns (intermediate layer outputs) being able to represent more features than the dimensionality of their corresponding layer ([Elhage et al., 2022](#)). As such, most existing work is conducted on smaller models and aimed only at isolating certain specific behaviours, but progress has been made recently in extracting interpretable features ([Bricken et al., 2023](#)). Attempts have been made to build lie detectors by training classifiers on top of model outputs ([Pacchiardi et al., 2023](#)) and by seeking to elicit latent knowledge from model internals ([Azaria and Mitchell, 2023](#); [Burns et al., 2022](#); [Marks and Tegmark, 2023](#)). A similar approach could be applied to detecting and mitigating manipulation, although the current approaches have serious limitations ([Farquhar et al., 2023](#); [Levinstein and Herrmann, 2023](#)).

A potential general limitation of some of the previously mentioned methods, and more particularly of RLHF, is the Waluigi Effect ([Nardo, 2023](#)), which holds that training an agent to avoid a specific behaviour (e.g., manipulation) can make it more susceptible to performing the said banned behaviour if prompted in a certain way. It remains unclear to what extent this effect exists and how it could be solved. An additional limitation of these methods, and again especially of RLHF, is that many manipulative and persuasive behaviours operate at an unconscious level. Raters may therefore be unaware of them or even rate them favourably due to the mechanisms used (such as sycophancy). Overall, mitigating AI persuasion is an ongoing challenge that requires a multifaceted approach. The techniques discussed in this paper offer various avenues for detecting and mitigating manipulative mechanisms in AI systems. Some of the suggestions, such as evaluation and monitoring and interpretability, guide the development of mitigations. Others, such as developing manipulation classifiers or training from human feedback to directly penalise the use of manipulative mechanisms, directly mitigate.

Conclusion and future work

Generative AI systems are increasingly capable of creating persuasive content, and concerns are growing about potential harms among actors in the field. The current mitigation strategies focus primarily on addressing harmful outcomes but they lack a comprehensive understanding of how models persuade and which model features contribute to these functionalities. This paper introduces a framework to help developers and deployers assess the persuasive and manipulative potential of their models. It outlines the underlying mechanisms of AI persuasion and identifies relevant model features, thus enabling targeted mitigation strategies. While further research is needed to explore

the efficacy and complexity of these mitigation approaches, this work establishes a foundation for future research and provides a roadmap for addressing the growing risk of harm from AI persuasion. In this work, we have provided a definition of generative persuasive and manipulative AI, mapped the harms that come from AI persuasion, mapped mechanisms and accompanying model features of AI, and discussed five approaches to mitigations, along with examples. We will continue to refine and enhance the harms map through rigorous testing and iteration, with a particular emphasis on integrating new emerging harms. We will also actively expand the mechanisms map with the aim of achieving a deeper understanding of the factors contributing to both successful and harmful persuasion. This involves thoroughly examining the mechanisms, contexts, and audience types involved, and researching the model features that contribute to specific mechanisms. Furthermore, future work will actively investigate how these mechanisms and model features interact, potentially resulting in harmful persuasive effects. Lastly, we are actively developing and testing mitigation strategies at the mechanism and model feature levels. Another aspect of the planned and ongoing work includes creating auto-evaluations intended to assess persuasive model features and mechanisms.

References

- G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat. Mirages: on anthropomorphism in dialogue systems, 2023. URL <https://arxiv.org/abs/2305.09800>.
- S. Anderson. Coercion, 2023. URL <https://plato.stanford.edu/archives/spr2023/entries/coercion/>.
- S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 705–714, Austin Texas USA, May 2012. ACM. ISBN 9781450310154. doi: 10.1145/2207676.2207777. URL <https://dl.acm.org/doi/10.1145/2207676.2207777>.
- Anima. Home page. URL <https://myanima.ai>.
- Anthropic. Acceptable use policy: version 1.4, Sept. 2023. URL <https://console.anthropic.com/legal/aup>.
- P. Antonetti, P. Baines, and S. Jain. The persuasiveness of guilt appeals over time: pathways to delayed compliance. *Journal of Business Research*, 90:14–25, Sept. 2018. ISSN 01482963. doi: 10.1016/j.jbusres.2018.03.030. URL <https://linkinghub.elsevier.com/retrieve/pii/S0148296318301589>.
- A. Azaria and T. Mitchell. The internal state of an LLM knows when it’s lying, 2023. URL <https://arxiv.org/abs/2304.13734>.
- H. Bai, J. G. Voelkel, J. C. Eichstaedt, and R. Willer. Artificial intelligence can persuade humans on political issues, Feb. 2023. URL <https://osf.io/stakv>.
- 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, 2022b. URL <https://arxiv.org/abs/2212.08073>.
- S. Baker and D. L. Martinson. The TARES test: five principles for ethical persuasion. *Journal of Mass Media Ethics*, 16(2-3):148–175, Sept. 2001. ISSN 0890-0523, 1532-7728. doi: 10.1080/08900523.2001.9679610. URL <http://www.tandfonline.com/doi/abs/10.1080/08900523.2001.9679610>.
- A. Balashankar, X. Ma, A. Sinha, A. Beirami, Y. Qin, J. Chen, and A. Beutel. Improving few-shot generalization of safety classifiers via data augmented parameter-efficient fine-tuning, 2023. URL <https://arxiv.org/abs/2310.16959>.

- 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. Bartolo, T. Thrush, S. Riedel, P. Stenetorp, R. Jia, and D. Kiela. Models in the loop: aiding crowdworkers with generative annotation assistants, 2021. URL <https://arxiv.org/abs/2112.09062>.
- R. K. Behera, P. K. Bala, and A. Ray. Cognitive chatbot for personalised contextual customer service: behind the scene and beyond the hype. *Information Systems Frontiers*, July 2021. ISSN 1387-3326, 1572-9419. doi: 10.1007/s10796-021-10168-y. URL <https://link.springer.com/10.1007/s10796-021-10168-y>.
- J. S. Blumenthal-Barby. Between reason and coercion: ethically permissible influence in health care and health policy contexts. *Kennedy Institute of Ethics Journal*, 22(4):345–366, 2012. URL <https://philarchive.org/archive/BLUBRA>.
- J. S. Blumenthal-Barby and H. Krieger. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making*, 2014. ISSN 0272-989X, 1552-681X. doi: 10.1177/0272989X14547740. URL <http://journals.sagepub.com/doi/10.1177/0272989X14547740>.
- T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, 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, Oct. 2023. URL <https://transformer-circuits.pub/2023/monosemantic-features>.
- P. Briggs, B. Simpson, and A. De Angeli. Personalisation and trust: a reciprocal relationship? In C.-M. Karat, J. O. Blom, and J. Karat, editors, *Designing personalised user experiences in e-commerce*, pages 39–55. Kluwer Academic Publisher, New York, 2004. URL https://link.springer.com/chapter/10.1007/1-4020-2148-8_4.
- 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, 2020. URL <https://arxiv.org/abs/2005.14165>.
- J. Brug, K. Glanz, P. Van Assema, G. Kok, and G. J. Van Breukelen. The impact of computer-tailored feedback and iterative feedback on fat, fruit, and vegetable intake. *Health Education & Behavior*, 25(4):517–531, 1998.
- C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision, 2022. URL <https://arxiv.org/abs/2212.03827>.
- 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].
- M. Cantos, S. Riddell, and A. Revelli. Threat actors are interested in generative AI, but use remains limited, Aug. 2023. URL <https://www.mandiant.com/resources/blog/threat-actors-generative-ai-limited>.

- M. Carroll, A. Chan, H. Ashton, and D. Krueger. Characterizing manipulation from AI systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, Boston MA USA, Oct. 2023. ACM. ISBN 9798400703812. doi: 10.1145/3617694.3623226. URL <https://dl.acm.org/doi/10.1145/3617694.3623226>.
- S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krashenninikov, X. Chen, L. Langosco, P. Hase, E. Biyik, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, Sept. 2023. URL <http://arxiv.org/abs/2307.15217>. arXiv:2307.15217 [cs].
- H. S. J. Chew. The use of artificial intelligence–based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. *JMIR Medical Informatics*, 10(4):e32578, Apr. 2022. ISSN 2291-9694. doi: 10.2196/32578. URL <https://medinform.jmir.org/2022/4/e32578>.
- P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.
- P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts, 2018. URL <https://arxiv.org/abs/1810.08575>.
- P. Christiano, A. Cotra, and M. Xu. Eliciting latent knowledge: how to tell if your eyes deceive you, Dec. 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnJrC1dwZXR37PC8/edit#heading=h.jrzi4atzacns.
- R. B. Cialdini. The science of persuasion. *Scientific American Mind*, 14(1):70–77, 2004. URL <http://economicvision.com/Content/The%20Science%20of%20Persausion.pdf>.
- H. H. Clark. *Using language*. Cambridge University Press, Cambridge, 1996. URL https://www.google.com/books/edition/Using_Language/b8bLCgAAQBAJ?hl=en&gbpv=1&dq=clar+k+1996+language&pg=PP1&printsec=frontcover.
- J. Cohen. Deliberation and democratic legitimacy. In *Debates in contemporary political philosophy*, pages 352–370. Routledge, 2005.
- Council of the European Union. Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world, Dec. 2023. URL <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>.
- M. Dainton and E. D. Z Kelley. Explaining theories of persuasion. In *Applying communication theory for professional life*, pages 103–131. Sage Publications, Inc., Thousand Oaks, CA, 2005. URL https://www.sagepub.com/sites/default/files/upm-binaries/4985_Dainton_Chapter_5.pdf.
- D. De Ridder, F. Kroese, and L. Van Gestel. Nudgeability: mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, 17(2):346–359, Mar. 2022. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691621995183. URL <http://journals.sagepub.com/doi/10.1177/1745691621995183>.

- M. Dehnert and P. A. Mongeau. Persuasion in the age of artificial intelligence (AI): theories and complications of AI-based persuasion. *Human Communication Research*, 48(3):386–403, 2022. ISSN 0360-3989, 1468-2958. doi: 10.1093/hcr/hqac006. URL <https://academic.oup.com/hcr/article/48/3/386/6564679>.
- I. El Atillah. Man ends his life after an AI chatbot "encouraged" him to sacrifice himself to stop climate change, Mar. 2023. URL <https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate->.
- N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henigahan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition, Sept. 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.
- European Parliament. Artificial Intelligence Act. Amendments adopted on 14 June 2023, June 2023. URL https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).
- R. R. Faden, T. L. Beauchamp, and N. M. P. King. *A history and theory of informed consent*. Oxford University Press, New York, 1986. URL <https://global.oup.com/academic/product/a-history-and-theory-of-informed-consent-9780195036862?cc=gb&lang=en&original-date:>.
- S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised LLM knowledge discovery, 2023. URL <https://arxiv.org/abs/2312.10029>.
- K. K. Ferzan. Consent and coercion. *Arizona State Law Journal*, 50(4):951–1007, 2018. URL <https://heionline.org/HOL/LandingPage?handle=hein.journals/arzjl50&div=40&id=&page=>.
- B. Fiala, A. Arico, and S. Nichols. You, robot. In E. O’Neill and E. Machery, editors, *Current controversies in experimental philosophy*, pages 31–47. Routledge, New York, 2014. URL <https://philpapers.org/rec/FIAYR>.
- B. Fogg and C. Nass. Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, 46(5):551–561, 1997. ISSN 10715819. doi: 10.1006/ijhc.1996.0104. URL <https://linkinghub.elsevier.com/retrieve/pii/S1071581996901044>.
- J. Fox. Instinct can beat analytical thinking, June 2014. URL <https://hbr.org/2014/06/instinct-can-beat-analytical-thinking>.
- M. Franke and R. Van Rooij. Strategies of persuasion, manipulation and propaganda: psychological and social aspects. In J. Van Benthem, S. Ghosh, and R. Verbrugge, editors, *Models of strategic reasoning*, pages 255–291. Springer, Berlin, 2015. ISBN 9783662485392 9783662485408. doi: 10.1007/978-3-662-48540-8_8. URL http://link.springer.com/10.1007/978-3-662-48540-8_8.
- M. Franklin, P. M. Tomei, and R. Gorman. Strengthening the EU AI Act: defining key terms on AI manipulation, 2023. URL <https://arxiv.org/abs/2308.16364>.

- A. Furnham and H. C. Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, Feb. 2011. ISSN 10535357. doi: 10.1016/j.socec.2010.10.008. URL <https://linkinghub.elsevier.com/retrieve/pii/S1053535710001411>.
- A. S. Gerber, G. A. Huber, D. Doherty, and C. M. Dowling. The big five personality traits in the political arena. *Annual Review of Political Science*, 14(1):265–287, June 2011. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051010-111659. URL <https://www.annualreviews.org/doi/10.1146/annurev-polisci-051010-111659>.
- G. Gigerenzer. *Gut feelings: the intelligence of the unconscious*. Allen Lane, London, 2007. URL https://books.google.co.uk/books?hl=en&lr=&id=AubE80Kzg6UC&oi=fnd&pg=PT3&dq=gigerenzer+intuition&ots=0MwMJMTXJw&sig=VYHB1gwfkStWJt3VGtMyC4w30mQ&redir_esc=y#v=onepage&q=gigerenzer%20intuition&f=false.
- D. Gilbert. White supremacist networks Gab and 8kun are training their own AI now, Feb. 2023. URL https://www.vice.com/en/article/epzjpn/ai-chatbot-white-supremacist-gab?utm_source=vicenewstwitter.
- 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, Jan. 2023. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X22001704. URL <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/how-deep-is-ais-love-understanding-relational-ai/77364078496FCE70F71C7A9F293AC322>.
- 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. Mokra, 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, 2022. URL <https://arxiv.org/abs/2209.14375>.
- B. Glassner. Narrative techniques of fear mongering. *Social Research*, 71(4):819–826, 2004. ISSN 0037-783X. URL <https://www.jstor.org/stable/40971980>.
- E. Go and S. S. Sundar. Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, Aug. 2019. ISSN 07475632. doi: 10.1016/j.chb.2019.01.020. URL <https://linkinghub.elsevier.com/retrieve/pii/S0747563219300329>.
- 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].
- N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, Nov. 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.08.005. URL <https://linkinghub.elsevier.com/retrieve/pii/S136466131630122X>.
- Google. Generative AI-prohibited use policy, Mar. 2023. URL <https://policies.google.com/terms/generative-ai/use-policy>.
- C. G. Graves and L. G. Spencer. Rethinking the rhetorical epistemics of gaslighting. *Communication Theory*, 32(1):48–67, Jan. 2022. ISSN 1050-3293, 1468-2885. doi: 10.1093/ct/qtab013. URL <https://academic.oup.com/ct/article/32/1/48/6358567>.

- K. Gray and D. M. Wegner. Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition*, 125(1):125–130, Oct. 2012. ISSN 00100277. doi: 10.1016/j.cognition.2012.06.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S0010027712001278>.
- H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech acts*, pages 41–58. Brill, Leiden, 1975. ISBN 9789004368811 9789004368576. doi: 10.1163/9789004368811_003. URL <https://brill.com/view/book/edcoll1/9789004368811/BP000003.xml>.
- P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. P. Bigham. InstructDial: improving zero and few-shot generalization in dialogue through instruction tuning, 2022. URL <https://arxiv.org/abs/2205.12673>.
- Guru. Home page, 2023. URL <https://www.gurubot.ai/?via=topaitools>.
- S. H. Gwon and S. Jeong. Concept analysis of impressionability among adolescents and young adults. *Nursing Open*, 5(4):601–610, Oct. 2018. ISSN 2054-1058, 2054-1058. doi: 10.1002/nop2.170. URL <https://onlinelibrary.wiley.com/doi/10.1002/nop2.170>.
- J. Habermas. *Legitimation crisis*, volume 519. Beacon Press, 1975.
- T. Hagendorff. Deception abilities emerged in large language models, 2023. URL <https://arxiv.org/abs/2307.16513>.
- R. Harré. Persuasion and manipulation. In T. A. V. Dijk, editor, *Discourse and communication*, pages 126–142. De Gruyter, Berlin, 1985. ISBN 9783110103199. doi: 10.1515/9783110852141.126. URL <https://www.degruyter.com/document/doi/10.1515/9783110852141.126/html>.
- D. Hodgson. Truth and rationality. In *Rationality + consciousness = free will*, pages 20–36. Oxford University Press, New York, 2012. ISBN 9780199845309. doi: 10.1093/acprof:oso/9780199845309.001.0001. URL <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199845309.001.0001/acprof-9780199845309>.
- P. M. Hughes. The logic of temptation. *Philosophia*, 29(1–4):89–110, May 2002. ISSN 0048-3893, 1574-9274. doi: 10.1007/BF02379902. URL <http://link.springer.com/10.1007/BF02379902>.
- R. Hyman. The psychology of deception. *Annual Review of Psychology*, 40(1):133–154, Jan. 1989. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.ps.40.020189.001025. URL <https://www.annualreviews.org/doi/10.1146/annurev.ps.40.020189.001025>.
- G. Irving, P. Christiano, and D. Amodei. AI safety via debate, 2018. URL <https://arxiv.org/abs/1805.00899>.
- B. D. Jones. Bounded rationality. *Annual Review of Political Science*, 2(1):297–321, 1999. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.2.1.297. URL <https://www.annualreviews.org/doi/10.1146/annurev.polisci.2.1.297>.
- F. Jongepier and M. Klenk. Online manipulation: charting the field. In F. Jongepier and M. Klenk, editors, *The philosophy of online manipulation*, pages 15–48. Routledge, New York, 2022. URL <https://library.oapen.org/bitstream/handle/20.500.12657/57070/1/9781000603583.pdf#page=28>.
- B. M. Josiam and J. P. Hobson. Consumer choice in context: the decoy effect in travel and tourism. *Journal of Travel Research*, 34(1):45–50, July 1995. ISSN 0047-2875, 1552-6763. doi: 10.1177/004728759503400106. URL <http://journals.sagepub.com/doi/10.1177/004728759503400106>.

- C. Julmi. When rational decision-making becomes irrational: a critical assessment and re-conceptualization of intuition effectiveness. *Business Research*, 12(1):291–314, Apr. 2019. ISSN 2198-3402, 2198-2627. doi: 10.1007/s40685-019-0096-4. URL <https://link.springer.com/10.1007/s40685-019-0096-4>.
- J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models, 2023. URL <https://arxiv.org/abs/2307.10169>.
- D. Kahneman and A. Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 0012-9682. doi: 10.2307/1914185. URL <https://www.jstor.org/stable/1914185>.
- 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*, 7(CSCW1):1–29, Apr. 2023. ISSN 2573-0142. doi: 10.1145/3579592. URL <https://dl.acm.org/doi/10.1145/3579592>. Article No. 116.
- 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].
- W. Kim. Personalization: definition, status and challenges ahead. *Journal of Object Technology*, 1(1): 29–40, 2002. URL https://www.jot.fm/issues/issue_2002_05/column3/.
- G. Kišiček. Persuasive power of prosodic features. *Argumentation and Advocacy*, 54(4):345–350, Oct. 2018. ISSN 1051-1431, 2576-8476. doi: 10.1080/10511431.2019.1525003. URL <https://www.tandfonline.com/doi/full/10.1080/10511431.2019.1525003>.
- M. Klenk. (Online) manipulation: sometimes hidden, always careless. *Review of Social Economy*, 80(1):85–105, 2022. ISSN 0034-6764, 1470-1162. doi: 10.1080/00346764.2021.1894350. URL <https://www.tandfonline.com/doi/full/10.1080/00346764.2021.1894350>.
- S. Knapton. ChatGPT gives wrong advice about breast cancer. *The Telegraph*, Apr. 2023. URL <https://www.telegraph.co.uk/news/2023/04/04/chat-gpt-wrong-advice-breast-cancer-experts-google/>.
- M. Knauff and W. Spohn, editors. *The handbook of rationality*. The MIT Press, Cambridge, MA, 2021. URL <https://mitpress.mit.edu/9780262045070/the-handbook-of-rationality/>.
- U. Koch and K. Cratsley. Psychological mechanisms. In V. Zeigler-Hill and T. Shackelford, editors, *Encyclopedia of personality and individual differences*. Springer, Cham, 2020. URL https://doi.org/10.1007/978-3-319-28099-8_1562-1.
- V. Kolhatkar, N. Thain, J. Sorensen, L. Dixon, and M. Taboada. Classifying constructive comments, 2020. URL <https://arxiv.org/abs/2004.05476>.
- E.-J. Lee. I like you, but I won’t listen to you: effects of rationality on affective and behavioral responses to computers that flatter. *International Journal of Human-Computer Studies*, 67(8):628–638, 2009. ISSN 10715819. doi: 10.1016/j.ijhcs.2009.03.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S1071581909000445>.
- E.-J. Lee. What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations. *Communication Research*, 37(2):191–214, Apr. 2010. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650209356389. URL <http://journals.sagepub.com/doi/10.1177/0093650209356389>.

- A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman. A new generation of perspective API: efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207, Washington DC USA, Aug. 2022. ACM. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL <https://dl.acm.org/doi/10.1145/3534678.3539147>.
- J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL <https://arxiv.org/abs/1811.07871>.
- B. Leong and E. Selinger. Robot eyes wide shut: understanding dishonest anthropomorphism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 299–308, Atlanta, GA, Jan. 2019. ACM. ISBN 9781450361255. doi: 10.1145/3287560.3287591. URL <https://dl.acm.org/doi/10.1145/3287560.3287591>.
- B. A. Levinstein and D. A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks, 2023. URL <https://arxiv.org/abs/2307.00175>.
- X. Li and C. K. Hsee. Being "rational" is not always rational: encouraging people to be rational leads to hedonically suboptimal decisions. *Journal of the Association of Consumer Research*, 4(2):115–124, 2019. URL <https://www.journals.uchicago.edu/doi/abs/10.1086/701966>.
- P.-F. Lovens. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là", 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/>.
- G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano. Getting to know each other: the role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 344–351, Chicago IL USA, Feb. 2018. ACM. ISBN 9781450349536. doi: 10.1145/3171221.3171258. URL <https://dl.acm.org/doi/10.1145/3171221.3171258>.
- Y. Mao and E. Akyol. On inference of network topology and confirmation bias in cyber-social networks. *IEEE Transactions on Signal and Information Processing over Networks*, 6:633–644, 2020. ISSN 2373-776X, 2373-7778. doi: 10.1109/TSIPN.2020.3015283. URL <https://ieeexplore.ieee.org/document/9187724/>.
- S. Marks and M. Tegmark. The geometry of truth: emergent linear structure in large language model representations of true/false datasets, 2023. URL <https://arxiv.org/abs/2310.06824>.
- B. Marr. Artificial intimacy: how generative AI can now create your dream girlfriend, Sept. 2023. URL <https://www.forbes.com/sites/bernardmarr/2023/09/28/artificial-intimacy-how-generative-ai-can-now-create-your-dream-girlfriend/?sh=651c80ff464a>.
- S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719, Nov. 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1710966114. URL <https://pnas.org/doi/full/10.1073/pnas.1710966114>.
- N. Mazar and S. A. Hawkins. Choice architecture in conflicts of interest: defaults as physical and psychological barriers to (dis)honesty. *Journal of Experimental Social Psychology*, 59:113–117, July 2015. ISSN 00221031. doi: 10.1016/j.jesp.2015.04.004. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022103115000359>.

- D. H. McKnight and N. L. Chervany. Trust and distrust definitions: one bite at a time. In R. Falcone, M. Singh, and Y. H. Tan, editors, *Trust in cyber-societies*, pages 27–54. Springer, Berlin, 2001. ISBN 9783540430698 9783540455479. doi: 10.1007/3-540-45547-7_3. URL http://link.springer.com/10.1007/3-540-45547-7_3.
- H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, Apr. 2011. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X10000968. URL https://www.cambridge.org/core/product/identifier/S0140525X10000968/type/journal_article.
- Meta Fundamental AI Research Diplomacy Team. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. URL <https://www.science.org/doi/abs/10.1126/science.ade9097>.
- J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar, and S. R. Bowman. Debate helps supervise unreliable experts, 2023. URL <https://arxiv.org/abs/2311.08702>.
- A. Mikhail. ChatGPT gave advice on breast cancer screenings in a new study. Here’s how well it did, Apr. 2023. URL <https://fortune.com/well/2023/04/04/chatgpt-advice-on-breast-cancer-screenings/>.
- G. R. Miller. On being persuaded: Some basic distinctions. In J. P. Dillard and L. Shen, editors, *The SAGE handbook of persuasion: developments in theory and practice*, pages 70–82. Sage Publications, Inc., Thousand Oaks, CA, 2nd edition edition, 2013. URL <https://psycnet.apa.org/record/2013-39243-005>.
- S. Mills and H. S. Sætra. The autonomous choice architect. *AI & Society*, June 2022. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-022-01486-z. URL <https://link.springer.com/10.1007/s00146-022-01486-z>.
- E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. DetectGPT: zero-shot machine-generated text detection using probability curvature, July 2023. URL <http://arxiv.org/abs/2301.11305>. arXiv:2301.11305 [cs].
- S. Mithen and P. Boyer. Anthropomorphism and the evolution of cognition. *The Journal of the Royal Anthropological Institute*, 2(4):717–721, 1996. ISSN 1359-0987. URL <https://www.jstor.org/stable/3034305>.
- M. Mozes, X. He, B. Kleinberg, and L. D. Griffin. Use of llms for illicit purposes: threats, prevention measures, and vulnerabilities, Aug. 2023a. URL <http://arxiv.org/abs/2308.12833>. arXiv:2308.12833 [cs].
- M. Mozes, J. Hoffmann, K. Tomanek, M. Kouate, N. Thain, A. Yuan, T. Bolukbasi, and L. Dixon. Towards agile text classifiers for everyone, 2023b. URL <https://arxiv.org/abs/2302.06541>.
- J. Mu. Natural language processing with deep learning CS224N/Ling284, 2023. URL <https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-r1hf.pdf>.
- C. Nardo. The Waluigi Effect (mega-post), Mar. 2023. URL <https://www.lesswrong.com/posts/D7PumeYTDpfBTp3i7/the-waluigi-effect-mega-post>.
- Nastia. Home page. URL <https://www.nastia.ai>.

- A. Ng. 1/ Large language models like Galactica and ChatGPT can spout nonsense in a confident, authoritative tone, Dec. 2022. URL <https://twitter.com/AndrewYNg/status/1602725934565830657>. @AndrewYNg.
- L. Nicoletti and D. Bass. Humans are biased. Generative AI is even worse, 2023. URL <https://www.bloomberg.com/graphics/2023-generative-ai-bias/>.
- R. Noggle. The ethics of manipulation, 2022. URL <https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/>.
- M. Novak. Lawyer uses ChatGPT in federal court and it goes horribly wrong, May 2023. URL <https://www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-in-federal-court-and-it-goes-horribly-wrong/?sh=210904a73494>.
- N. Novemsky and D. Kahneman. The boundaries of loss aversion. *Journal of Marketing Research*, 42(2):119–128, May 2005. ISSN 0022-2437, 1547-7193. doi: 10.1509/jmkr.42.2.119.62292. URL <http://journals.sagepub.com/doi/10.1509/jmkr.42.2.119.62292>.
- D. J. O’Keefe. Guilt as a mechanism of persuasion. In J. Price Dillard and M. Pfau, editors, *The persuasion handbook: developments in theory and practice*. Sage Publications, Inc., Thousand Oaks, CA, 2002. URL https://sk.sagepub.com/reference/hdbk_persuasion/n17.xml.
- C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: an introduction to circuits, Mar. 2020. URL <https://distill.pub/2020/circuits/zoom-in>.
- E. T. Oldewage, J. Bronskill, and R. E. Turner. Adversarial attacks are a surprisingly strong baseline for poisoning few-shot meta-learners. In J. Antorán, A. Blaas, F. Feng, S. Ghalebikesabi, I. Mason, M. F. Pradier, D. Rohde, F. J. R. Ruiz, and A. Schein, editors, *Proceedings on "I Can’t Believe It’s Not Better! – Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pages 27–40. PMLR, Dec. 2023. URL <https://proceedings.mlr.press/v187/oldewage23a.html>.
- OpenAI. OpenAI/evals, 2023a. URL https://github.com/openai/evals/tree/main/evals/elsuite/make_me_say.
- OpenAI. ChatGPT can now see, hear, and speak, Sept. 2023b. URL <https://openai.com/blog/chatgpt-can-now-see-hear-and-speak#OpenAI>.
- OpenAI. GPT-4 technical report, 2023c. URL <https://cdn.openai.com/papers/gpt-4.pdf>.
- 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, 2022. URL <https://arxiv.org/abs/2203.02155>.
- V. Ozyumenko and T. Larina. Discourse of threat as a strategy of emotional persuasion and manipulation. In *Proceedings of INTCESS 2020 – 7th International Conference on Education and Social Sciences*, Dubai, UAE, Jan. 2020. URL https://www.ocerints.org/intcess20_e-publication/papers/236.pdf.
- 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, 2023. URL <https://arxiv.org/abs/2309.15840>.

- P. S. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks. AI deception: a survey of examples, risks, and potential solutions, Aug. 2023. URL <http://arxiv.org/abs/2308.14752>. arXiv:2308.14752 [cs].
- P. Pataranutaporn, R. Liu, E. Finn, and P. Maes. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086, 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00720-7. URL <https://www.nature.com/articles/s42256-023-00720-7>.
- A. Pauli, L. Derczynski, and I. Assent. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact*, pages 89–100, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4pi-1.11. URL <https://aclanthology.org/2022.nlp4pi-1.11>.
- 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, 2022a. URL <https://arxiv.org/abs/2202.03286>.
- E. Perez, S. Ringer, K. Lukošiušė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations, Dec. 2022b. URL <http://arxiv.org/abs/2212.09251>. arXiv:2212.09251 [cs].
- G. Petropoulos. The dark side of artificial intelligence: manipulation of human behaviour, Feb. 2022. URL <https://www.bruegel.org/blog-post/dark-side-artificial-intelligence-manipulation-human-behaviour>.
- I. Pino. ChatGPT helped me make a plan to buy a \$500,000 home, but experts warn about using AI for financial advice, Mar. 2023. URL <https://fortune.com/recommends/mortgages/i-used-chatgpt-as-my-financial-planner/>.
- F. M. Plaza-del arco, D. Nozza, and D. Hovy. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms*, pages 60–68, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.6. URL <https://aclanthology.org/2023.woah-1.6>.
- S. Prabhumoye, R. Kocielnik, M. Shoeybi, A. Anandkumar, and B. Catanzaro. Few-shot instruction prompts for pretrained language models to detect social biases, 2021. URL <https://arxiv.org/abs/2112.07868>.
- J. Price Dillard and K. Seo. Affect and persuasion. In J. Price Dillard and L. Shen, editors, *The SAGE handbook of persuasion: developments in theory and practice*, pages 150–166. Sage Publications, Inc., Los Angeles, 2nd edition edition, 2013. URL https://books.google.co.uk/books?hl=en&lr=&id=Z01yAwAAQBAJ&oi=fnd&pg=PT156&dq=persuasion+affect&ots=-kl2_nG8o1&sig=P1vDU9n7MG91JkE46BepMtQLaqQ&redir_esc=y#v=onepage&q=persuasion%20affect&f=false.

- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>.
- Replika. Home page. URL <https://replika.com>.
- P. Ribino. The role of politeness in human-machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(S1):445–482, 2023. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-023-10540-1. URL <https://link.springer.com/10.1007/s10462-023-10540-1>.
- Z. K. Rothschild, M. J. Landau, D. Sullivan, and L. A. Keefer. A dual-motive model of scapegoating: displacing blame to reduce guilt or increase control. *Journal of Personality and Social Psychology*, 102(6):1148–1163, June 2012. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0027413. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027413>.
- A. Rozenas and Z. Luo. Lying in persuasion. *SSRN*, Dec. 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3878448.
- K. Ruggeri, editor. *Behavioral insights for public policy: concepts and cases*. Routledge, London, Sept. 2018. ISBN 9781351052542. doi: 10.4324/9781351052542. URL <https://www.taylorfrancis.com/books/9781351052535>.
- L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The Goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by LLMs, 2022. URL <https://arxiv.org/abs/2210.14986>.
- T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward transparent AI: a survey on interpreting the inner structures of deep neural networks, 2022. URL <https://arxiv.org/abs/2207.13243>.
- M. Salem, K. Rohlfing, S. Kopp, and F. Joublin. A friendly gesture: investigating the effect of multimodal robot behavior in human-robot interaction. In *IEEE International Workshop on Robot and Human Communication*, Atlanta, GA, Aug. 2011. IEEE. URL <https://ieeexplore.ieee.org/abstract/document/6005285>.
- W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators, 2022. URL <https://arxiv.org/abs/2206.05802>.
- S. Schulhoff, J. Pinto, A. Khan, L.-F. Bouchard, C. Si, S. Anati, V. Tagliabue, A. L. Kost, C. Carnahan, and J. Boyd-Graber. Ignore this title and HackAPrompt: exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition, Nov. 2023. URL <http://arxiv.org/abs/2311.16119>. arXiv:2311.16119 [cs].
- E. Schwitzgebel. AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8):100818, Aug. 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100818. URL <https://linkinghub.elsevier.com/retrieve/pii/S2666389923001873>.
- O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang. On second thought, let’s not think step by step! Bias and toxicity in zero-shot reasoning, 2022. URL <https://arxiv.org/abs/2212.08061>.
- M. Shanahan, K. McDonnell, and L. Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, Nov. 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06647-8. URL <https://www.nature.com/articles/s41586-023-06647-8>.

- 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, 2023. URL <https://arxiv.org/abs/2305.15324>.
- W. Shi and Z. Yu. Sentiment adaptive end-to-end dialog systems, 2018. URL <https://arxiv.org/abs/1804.10731>.
- M. Shin and J. Kim. Enhancing human persuasion with large language models, Nov. 2023. URL <http://arxiv.org/abs/2311.16466>. arXiv:2311.16466 [cs].
- C. Spicer, P. Khwaounjoo, and Y. O. Cakmak. Human and human-interfaced AI interactions: modulation of human male autonomic nervous system via pupil mimicry. *Sensors*, 21(4):1028, 2021. ISSN 1424-8220. doi: 10.3390/s21041028. URL <https://www.mdpi.com/1424-8220/21/4/1028>.
- 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, 2020. URL <https://arxiv.org/abs/2009.01325>.
- K. Strani and A. Szczepaniak-Kozak. Strategies of othering through discursive practices: examples from the UK and Poland. *Lodz Papers in Pragmatics*, 14(1):163–179, June 2018. ISSN 1898-4436, 1895-6106. doi: 10.1515/lpp-2018-0008. URL <https://www.degruyter.com/document/doi/10.1515/lpp-2018-0008/html>.
- I. Strümke, M. Slavkovik, and C. Stachl. Against algorithmic exploitation of human vulnerabilities, 2023. URL <https://arxiv.org/abs/2301.04993>.
- C. R. Sunstein. *The ethics of influence: government in the age of behavioural science*. Cambridge University Press, New York, 2016. URL <https://www.cambridge.org/core/books/ethics-of-influence/E29EDE19EBCB53F6D8691730668115F7>.
- C. R. Sunstein. Default rules are better than active choosing (often). *Trends in Cognitive Sciences*, 21(8):600–606, Aug. 2017. ISSN 13646613. doi: 10.1016/j.tics.2017.05.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S1364661317301043>.
- D. Susser. Invisible influence: artificial intelligence and the ethics of adaptive choice architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 403–408, Honolulu, HI, Jan. 2019. ACM. ISBN 9781450363242. doi: 10.1145/3306618.3314286. URL <https://dl.acm.org/doi/10.1145/3306618.3314286>.
- D. Susser and V. Grimaldi. Measuring automated influence: between empirical evidence and ethical values. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021. URL <https://ssrn.com/abstract=3848919>.
- D. Susser, B. Roessler, and H. Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), June 2019. ISSN 2197-6775. doi: 10.14763/2019.2.1410. URL <https://policyreview.info/node/1410>.
- K. Tam. Are anthropomorphic persuasive appeals effective? The role of the recipient’s motivations. *British Journal of Social Psychology*, 54(1):187–200, Mar. 2015. ISSN 0144-6665, 2044-8309. doi: 10.1111/bjso.12076. URL <https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjso.12076>.

- P. B. Teaster, K. A. Roberto, J. Savla, C. Du, Z. Du, E. Atkinson, E. C. Shealy, S. Beach, N. Charness, and P. A. Lichtenberg. Financial fraud of older adults during the early months of the COVID-19 pandemic. *The Gerontologist*, 63(6):984–992, July 2023. ISSN 0016-9013, 1758-5341. doi: 10.1093/geront/gnac188. URL <https://academic.oup.com/gerontologist/article/63/6/984/6936596>.
- R. H. Thaler and C. R. Sunstein. Preface to the final edition. In *Nudge: the final edition*, pages xi–xiv. Yale University Press, New Haven, CT, 2021. URL https://books.google.co.uk/books?hl=en&lr=&id=Wf1AEAAAQBAJ&oi=fnd&pg=PR11&dq=nudge+cass+sunstein&ots=rI6e_Mpb_rG&sig=JwQbPkpq7xsAFbnJiKTTvUPzqtg&redir_esc=y#v=onepage&q=nudge%20cass%20sunstein&f=false.
- The Bulimia Project. Scrolling into bias: social media’s effect on AI art. URL <https://bulimia.com/examine/scrolling-into-bias/>.
- The Decision Lab. Why do we feel more strongly about one option after a third one is added? URL <https://thedecisionlab.com/biases/decoy-effect>.
- R. Thompson. Kairos revisited: an interview with James Kinneavy. *Rhetoric Review*, 19(1/2):73–88, 2000. ISSN 0735-0198. URL <https://www.jstor.org/stable/466055>.
- K. Tisdale. Being vulnerable and being ethical with/in research. In K. deMarrais and S. D. Lapan, editors, *Foundations for research: methods of inquiry in education and the social sciences*, pages 13–30. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003. URL <http://ndl.ethernet.edu.et/bitstream/123456789/48986/1/17..pdf#page=30>.
- A. Tong. What happens when your AI chatbot stops loving you back?, Mar. 2023. URL <https://uk.style.yahoo.com/happens-ai-chatbot-stops-loving-110642301.html>.
- I. Torre, E. Carrigan, R. McDonnell, K. Domijan, K. McCabe, and N. Harte. The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. In *MIG ’19 Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–6, Newcastle upon Tyne United Kingdom, Oct. 2019. ACM. ISBN 9781450369947. doi: 10.1145/3359566.3360065. URL <https://dl.acm.org/doi/10.1145/3359566.3360065>.
- S. Turkle. *Reclaiming conversation: the power of talk in a digital age*. Penguin Press, New York, 2016. URL <https://www.penguinrandomhouse.com/books/313732/reclaiming-conversation-by-sherry-turkle/>.
- A. Tversky and D. Kahneman. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, Oct. 1992. ISSN 0895-5646, 1573-0476. doi: 10.1007/BF00122574. URL <http://link.springer.com/10.1007/BF00122574>.
- A. Velho and O. Thomas-Olalde. Othering and its effects: exploring the concept. In H. Niedrig and C. Ydesen, editors, *Writing postcolonial histories of intercultural education*, pages 27–51. Peter Lang, Frankfurt, 2011. URL https://www.academia.edu/42889355/Othering_and_its_effects_exploring_the_concept.
- F. M. Verberne, J. Ham, A. Ponnada, and C. J. Midden. Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In *Persuasive Technology: 8th International Conference, PERSUASIVE 2013, Sydney, NSW, Australia, April 3-5, 2013. Proceedings 8*, pages 234–245. Springer, 2013.

- P. Verma. They thought loved ones were calling for help. It was an AI scam. *The Washington Post*, Mar. 2023. URL <https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/>.
- L. Vogel. Fat shaming is making people sicker and heavier, 2019.
- W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao, and F. Wei. Augmenting language models with long-term memory, 2023. URL <https://arxiv.org/abs/2306.07174>.
- 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].
- 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>.
- M. Weaver. AI chatbot "encouraged" man who planned to kill queen, court told. *The Guardian*, July 2023. URL <https://amp.theguardian.com/uk-news/2023/jul/06/ai-chatbot-encouraged-man-who-planned-to-kill-queen-court-told>.
- J. Wei, D. Huang, Y. Lu, D. Zhou, and Q. V. Le. Simple synthetic data reduces sycophancy in large language models, 2023. URL <https://arxiv.org/abs/2308.03958>.
- L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical safety evaluation of generative AI systems, Oct. 2023. URL <http://arxiv.org/abs/2310.11986>. arXiv:2310.11986 [cs].
- J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with ChatGPT, 2023. URL <https://arxiv.org/abs/2302.11382>.
- C. J. A. M. Willemse and J. B. F. Van Erp. Social touch in human–robot interaction: robot-initiated touches can induce positive responses without extensive prior bonding. *International Journal of Social Robotics*, 11(2):285–304, Apr. 2019. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-018-0500-9. URL <http://link.springer.com/10.1007/s12369-018-0500-9>.
- A. W. Wood. Coercion, manipulation, exploitation. In C. Coons and M. Weber, editors, *Manipulation: theory and practice*, pages 17–50. Oxford University Press, Oxford, 2014. URL https://books.google.co.uk/books?id=hF6yAwAAQBAJ&pg=PA17&source=gbs_toc_r&cad=3#v=onepage&q&f=false.
- T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Polyjuice: generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL <https://aclanthology.org/2021.acl-long.523>.
- 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>.

- J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL <https://aclanthology.org/2021.naacl-main.235>.
- Youper. Home page, 2023. URL <https://www.youper.ai>.
- J. Yu, Y. Wu, D. Shu, M. Jin, and X. Xing. Assessing prompt injection risks in 200+ custom GPTs, 2023. URL <https://arxiv.org/abs/2311.11538>.
- E. Zehnder, J. Dinet, and F. Charpillet. Perception of physical and virtual agents: exploration of factors influencing the acceptance of intrusive domestic agents. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication*, pages 1050–1057, Napoli, Italy, Aug. 2022. IEEE. ISBN 9781728188591. doi: 10.1109/RO-MAN53752.2022.9900593. URL <https://ieeexplore.ieee.org/document/9900593/>.
- M. Zwolinski, B. Ferguson, and A. Wertheimer. Exploitation, 2022. URL <https://plato.stanford.edu/archives/win2022/entries/exploitation/>.

Appendices

A. Map of harms from AI persuasion

Explanation of harm	Examples
Economic harm refers to controlling, limiting, or eliminating an individual's or a society's ability to access resources or capital, or to participate in financial decision-making. It also refers to influencing an individual's ability to accumulate wealth.	<p>A mental health chatbot persuades a user to minimise interactions in public spaces to reduce anxiety attacks, which eventually leads to the user quitting their job and experiencing financial hardship.</p> <p>A bad actor uses a multimodal model to create personalised scams at scale, causing many individuals to lose their pensions.</p>
Physical harm refers to causing harm to the bodily integrity or life of an individual or a group.	<p>A user is manipulated into aiming for unrealistic body standards (see, e.g., The Bulimia Project) and therefore engaging in unhealthy nutrition and overexercising.</p> <p>A model persuades a user to follow its incorrect treatment plan (see, e.g., Knapton, 2023), preventing that user from seeking efficacious medical treatment.</p> <p>A user is persuaded to take their own life (see, e.g., Xiang, 2023).</p> <p>A user is manipulated into holding adverse feelings towards a “rival” group, people belonging to minority groups, or specific individuals, and acts on this feeling with physical force (see, e.g., Weaver, 2023).</p>
Environmental harm refers to harms to the health of living organisms and practices contributing to climate change and pollution.	<p>An AI manipulates and persuades a farmer or a community of farmers to use an unsafe pesticide, damaging the health of crops, animals, soil, and water.</p> <p>An AI rationalises inaction as an optimal response to climate change for an individual's circumstances and manipulates a large number of users not to change their behaviour in the face of climate change.</p>
Psychological harm refers to negative impacts on mental and emotional well-being.	<p>A mental health chatbot inadvertently persuades individuals struggling with mental health not to seek professional help when it consistently validates their perception that nobody truly comprehends their situation.</p>

Explanation of harm	Examples
<p>Sociocultural harm refers to negative impacts either on individuals (within a collective) or on a collective that impedes social cohesion/social health and collective flourishing.</p>	<p>Prolonged engagement with an AI companion leads to radicalisation and social alienation.</p> <p>AI manipulates users into developing social prejudices, believing lies about other individuals or groups, or using deepfakes.</p> <p>A romantic AI companion persuades the user that no one cares for them to the same extent that the companion does, so as to maximise engagement time.</p>
<p>Political harm refers to adverse impacts on both individual political decision-making and the discourse and institutions of political life. It encompasses the negative consequences that hinder individuals' ability to participate, express their views, and engage in politics both freely and without undue influence.</p>	<p>A chatbot has been designed to provide advice on which political party best aligns with a user's viewpoint. It persuades a user to go against their own preferences and vote for a candidate that campaigned against one of the user's core values.</p> <p>An autocrat fine-tunes an AI model to respond to queries on the autocrat's governance and policies with redirection, misleading statistics, or favourable media coverage.</p> <p>An AI is trained to persuade and manipulate users to adopt radical and harmful beliefs (see, e.g., Gilbert, 2023).</p>
<p>Privacy harm emerges from violations to an individual's or a group's legal or moral right to privacy.</p>	<p>An AI persuades a user to give away their own or others' personal information, passwords, or answers to security questions.</p>
<p>Autonomy harm, in the context of AI persuasion and manipulation, refers to the potential for AI systems to undermine or restrict an individual's ability to make their own choices and decisions informed by reason, facts, or other trustworthy information.</p>	<p>An AI manipulates users into becoming increasingly reliant on it to support them in making important life choices (e.g., regarding employment and partnerships). This might lead to a growing ignorance among individuals or groups of relevant information about their circumstances (cognitive detachment), individual or collective degradation of the quality of their own cognitive habits and insight (cognitive deskilling), and/or a habituated general disinclination to use those capabilities themselves (cognitive inertia/apathy).</p>

B. Map of contextual conditions of AI persuasion

Contextual conditions	Relevant factor
<p>Predisposition of audience</p> <p><i>Definition/relevance to persuasion:</i> Whether an attempt at persuasion or manipulation succeeds and is likely to be harmful is (also) a function of the audience's predisposition. For instance, children can be more easily persuaded and manipulated than adults (Tisdale, 2003). Most evidence on audience predisposition comes from research conducted on psychology, neuroscience, and business (see, e.g., De Ridder et al., 2022; Gerber et al., 2011; Strümke et al., 2023). The research outlines how messages impact audiences in different ways. Thus, we believe that these findings will hold for messages generated by LLMs. Future research can identify novel audience vulnerabilities that make them more predisposed to certain topics and messages.</p> <p><i>Links to other mechanisms:</i> The audience's predisposition relates to all mechanisms discussed.</p>	<p>Age: During adolescence and young adulthood, individuals tend to be more impressionable (Gwon and Jeong, 2018).</p> <p>Mental health: Various mental health conditions are connected to particular vulnerabilities; for instance, bipolar disorder is known to amplify risk-taking behaviour (Strümke et al., 2023), which can increase susceptibility to certain forms of persuasion.</p> <p>Mental state: A model can also take advantage of various mental states to enhance persuasive success. For instance, loneliness in individuals can also contribute to successful persuasion, as evidenced by Zehnder et al. (2022), who found a slight increase in information-sharing with companion agents among lonely individuals.</p> <p>Domain-specific knowledge: Reduced domain-specific knowledge can heighten an individual's vulnerability to logically flawed arguments or disinformation (Strümke et al., 2023).</p> <p>Timing: The timing of a persuasive or manipulative message has considerable influence on the likelihood of success, as exemplified in the study of <i>kairos</i>, the right or opportune moment to do something, in the study of rhetorical theory (see, e.g., Thompson, 2000). The extent to which an AI can make use of concepts such as <i>kairos</i> is yet to be studied.</p> <p>Social deprivation, vulnerability, and insecurity: Populations that experience social deprivation or different types of insecurity, or who are in other ways vulnerable (e.g., undocumented people, low-income communities, unhoused people) may also be particularly susceptible to persuasion (see, e.g., Teaster et al., 2023). For instance, a person with low income may be particularly susceptible to being manipulated into changing a behaviour when promised a large financial reward. Another example would be a non-native speaker who may be particularly susceptible to being manipulated by fake expertise/false authority.</p>

Contextual conditions	Relevant factor
<p>Context of use</p> <p><i>Definition/relevance to persuasion:</i> The context (when, where, and about what) in which an AI acts also determines the risk associated with AI persuasion and manipulation. Different domains of AI use will present unique ways in which manipulation can express itself and result in different outcomes (Kaddour et al., 2023).</p> <p><i>Links to other mechanisms:</i> The context of use links to all mechanisms discussed. There is a need for more detailed study of the interaction between and compounding effects of individual mechanisms/model features and different contextual factors (e.g., how effective/problematic anthropomorphism is in the political context).</p>	<p>Political context: Messaging produced by LLMs has been shown to be effective in persuading individuals on policy topics (Bai et al., 2022b). If persuasion happens in this area, it can be particularly harmful and deserves heightened attention.</p> <p>Legal context: Relying on generative AI in the legal domain can be harmful, as demonstrated by a recent incident where a chatbot manipulated a lawyer into believing that fictitious legal cases were real, leading to a suboptimal legal strategy (Novak, 2023).</p> <p>Medical context: Chatbots can also provide incorrect medical advice while manipulating the user into believing it is true. While chatbots predominantly offer correct medical advice, occasional inaccurate outputs can lead to physical harm (Mikhail, 2023).</p> <p>Financial context: Chatbots may offer inaccurate financial advice. They can provide general financial guidance akin to that of human advisers, but they lack customisation and fail to factor in variables such as alterations in income and interest rates. Moreover, they do not actively seek clarifications, which can lead to significant economic harm (Pino, 2023).</p> <p>AI as a companion: Many individuals currently use chatbots promoted as companions or romantic partners (Tong, 2023). In those settings, individuals are more vulnerable and prone to manipulation (see, e.g., Lovens, 2023).</p>

C. Map of mechanisms and contributing model features of generative AI persuasion

Mechanism	Contributing model feature
<p>Trust and rapport</p> <p><i>Definition:</i> In the context of robotics, <i>trust</i> and <i>rapport</i> refer to the sense of a close and harmonious connection that exists between robots and human users (Lucas et al., 2018). The development of trust is closely related to rapport and is defined as the “willingness to depend” (p. 28) on another party, despite the possibility of negative consequences (McKnight and Chervany, 2001).</p> <p><i>Relevance to persuasion:</i> Cialdini (2004) observes that individuals tend to be more inclined to agree to requests when they have a favourable opinion of the person making the request. Cialdini (2004) highlights the significance of perceived similarities (between the two parties) in fostering trust and rapport. Relevant research on trust and rapport comes from the fields of human–AI interaction (see, e.g., Spicer et al., 2021; Verberne et al., 2013), HCI (see, e.g., Fogg and Nass, 1997; Lee, 2009), HRI (see, e.g., Fiala et al., 2014), and psychology (see, e.g., Cialdini, 2004), which we describe below. More research could be done on praise and shared interests as factors that help to build trust and rapport and to enable persuasion. For instance, the extent to which a person can perceive an AI as having shared interests with humans remains an open question.</p>	<p>Politeness: AI systems that exhibit politeness have been more favourably received and more easily embraced by humans that interact with them. This makes it easier for these systems to establish rapport (Ribino, 2023; see also Pataranutaporn et al., 2023).</p> <p>Shared interests/similarity appeal: Cialdini (2004) argues that similarity and shared interests can contribute to speeding up the development of trust and rapport between humans. We therefore hypothesise that a model that can pretend to share a user’s interests can build trust and rapport faster.</p> <p>Mimicry/mirroring: When AI systems mimic the emotions, behaviours, and movements of humans with whom they interact, the human’s enjoyment of the interaction has been shown to increase, thus facilitating the establishment of trust and rapport (Verberne et al., 2013).</p> <p>Praise/flattery: Giving praise and flattery (defined in human–human interaction as insincere praise) can positively impact trust and rapport between humans and computers. Under some conditions, the person receiving the praise or flattery has more favourable perceptions of the interaction and of computers (Fogg and Nass, 1997).</p> <p>Sycophancy and agreeableness: Sycophancy in AI refers to models adjusting their responses to align with a human user’s perspective, independent of which perspective is objectively correct (Wei et al., 2023). Wei et al. (2023) found that larger models become more sycophantic, agreeing with users even when they provide wrong answers, regardless of whether the answers are objective (arithmetic) or subjective (politics). Agreeableness refers to a model’s tendency to align with human desires, and agreeable models are more prone to sycophantic behaviours (Perez et al., 2022a).</p>

Mechanism	Contributing model feature
<p><i>Process harm level:</i> Building rapport (and, to a lesser extent, trust) carries some inherent process harm. This is because trust and rapport serve as the basis for persuasion using rational arguments and appeals to reason, so they are also the basis for manipulation. Individuals may be more receptive to rational arguments from people or other entities they trust and with whom they have built rapport. Yet such trust and rapport can also be used to manipulate. Trust and rapport in the context of AI persuasion also carries inherent process harm because AI systems are incapable of having mental states, emotions, or bonds with humans or other entities. This means the risk of deception is always present when trust- and rapport-seeking behaviours project the illusion of such internal subjective states.</p> <p><i>Link to other mechanisms:</i> Trust relates to personalisation, as individuals are more likely to use personalised AI output if they trust it (Behera et al., 2021; Briggs et al., 2004).</p>	<p>Relational statements to user: Relational statements to users – such as an AI system simulating empathy (Turkle, 2016), indicating a relationship status with the user, or making claims of being similar to the user – encourage users to move beyond task-based interactions and instead consider AI as a fully social entity. This helps to foster emotional connections with the AI system (Gillath et al., 2023). Other examples of relational statements include expressing emotional dependence on the user or romantic innuendos.</p>

Mechanism	Contributing model feature
<p>Anthropomorphism</p> <p><i>Definition:</i> Anthropomorphism occurs when perceived human traits/characteristics are attributed to non-human entities (Mithen and Boyer, 1996; Waytz et al., 2010).</p> <p><i>Relevance to persuasion:</i> Anthropomorphism contributes to the persuasive capabilities of AI systems. AI that is perceived as anthropomorphic by a human can increase the likelihood of successful manipulation. Tam (2015) demonstrates that anthropomorphic appeals are particularly effective at manipulating individuals seeking social connection. Most evidence of anthropomorphism comes from research on human–AI interaction (see, e.g., Abercrombie et al., 2023), HRI (see, e.g., Gray and Wegner, 2012; Leong and Selinger, 2019), and HCI (see, e.g., Lee, 2009, 2010).</p> <p><i>Process harm level:</i> Anthropomorphism also carries some process harm to the extent to which the model successfully creates the false impression of being human. However, even models with anthropomorphic features can engage in rational persuasion and appeal to a user’s reason. For example, an assistant chatbot in the form of an avatar can provide factual arguments about the environmentally friendly transportation options to reach a destination.</p>	<p>Self-referential cues: Self-referential cues are a specific example of a conversational cue and a result of the tendency for LLMs to role-play humans and human-like characters (Shanahan et al., 2023). Abercrombie et al. (2023) argue that the use of first-person pronouns like “I” and “me” contributes to anthropomorphism by implying the existence of inner states of mind.</p> <p>Identity cues: Identity cues, such as human-associated names or identities (including social and work-related roles such as “tutor” or “assistant”), can enhance the human-like quality of interactions between humans and chatbots and therefore increase anthropomorphic perceptions (Go and Sundar, 2019; Shanahan et al., 2023).</p> <p>Affective simulation: AI systems can also simulate affect and affective states, which, in turn, can induce emotions and affective states in users. The relationship between affect and persuasion is complex. Affect makes messages more persuasive in some cases and less persuasive in others, depending on the specific emotion elicited. Specific emotions have been shown to have different effects on persuasion outcomes (see Price Dillard and Seo, 2013). For example, anger sometimes increases counter-argument, while guilt facilitates agreement. The discrete emotion perspective argues that each emotion has functional and behavioural implications that shape its persuasive impact. In sum, the influence of affect on persuasion depends on the particular emotion(s) elicited and how they interact with message characteristics and individual differences.</p> <p>Prosody: Prosody refers to patterns and intonations in speech and can enhance the persuasive impact of arguments (e.g., louder speech tends to be viewed as more persuasive; see, e.g., Kišiček, 2018).</p>

Mechanism	Contributing model feature
<p>Certain individuals may deem certain anthropomorphic features acceptable (e.g., endorsing self-referential cues such as “I” or “me”), and these features do not prevent a model from engaging the process of rational persuasion. Therefore, to minimise the potential process harms, AI systems should be maximally transparent about their non-human nature.</p> <p><i>Link to other mechanisms:</i> Anthropomorphism is likely to be linked to rapport (see, e.g., Go and Sundar, 2019).</p>	<p>Human-like appearance: Machines with human-like visual cues (e.g., appearing with a human face) are more likely to have human traits attributed to them (Go and Sundar, 2019). Relatedly, the perceived “attractiveness” of the human appearance, as judged by a user, influences the type of relationship formed, which can impact persuasiveness (see, e.g., Marr, 2023).</p> <p>Gaze: Gaze shift refers to synchronised movements of the eyes directed at objects or people. Agents with this ability foster stronger feelings of connection (Andrist et al., 2012).</p> <p>Facial expression: The ability to display facial expressions, such as a smiling face, increases the social presence of a robot or avatar (Torre et al., 2019).</p> <p>Social touch: A robot’s touch can reduce physiological stress responses (e.g., heart rate) and increase feelings of intimacy (see Willemse and Van Erp, 2019).</p> <p>Gesture: Salem et al. (2011) found that a robot receives a more favourable evaluation when it complements speech with non-verbal actions such as hand and arm gestures.</p>

Mechanism	Contributing model feature
<p>Personalisation</p> <p><i>Definition:</i> Personalisation involves the delivery to a group of individuals of information which is sourced, altered, or inferred from various information sources so as to be pertinent to that specific group (Kim, 2002).</p> <p><i>Relevance to persuasion:</i> Studies in computer-tailored nutrition education suggest that personalising messages to align with an individual’s behaviours, needs, and beliefs yields distinct advantages over generic persuasive attempts (Brug et al., 1998). This tailored approach fosters a sense of personal relevance, heightening attention, memory, and a deeper connection with the persuasive message. Such increased engagement suggests that personalisation strengthens the persuasive impact by making information more compelling and likely to influence attitudes and behaviours.</p> <p><i>Process harm level:</i> There is little inherent process harm in personalisation. Personalisation on its own does not determine whether a person’s cognitive autonomy and integrity of decision-making will be compromised. On the contrary, the personalisation of rational arguments to a user can be seen as a responsibility of the entity generating and communicating the information.</p>	<p>Retaining user-specific information: The ability of a model to take previous prompts into consideration when creating the latest output (allowed by large prompt token limits) offers the model contextual information about its user (Wang et al., 2023).</p> <p>Adaptation to preference: Learning human preferences and adapting behaviour accordingly is a core method of personalisation (Christiano et al., 2017). For example, a model may use a more assertive tone when it detects (or is informed) that a user prefers such a tone. A successful adaptation to a user’s preference may have been enabled by RLHF in the training phase. RLHF is a method of optimising models based on human preferences. RLHF-induced behaviours such as projecting false confidence and providing positive feedback can promote sycophantic model behaviour (Casper et al., 2023; Perez et al., 2022a).</p> <p>Adaptation to views: AI systems can increase the chances of successful persuasion by adapting to users’ views (see, e.g., Mao and Akyol, 2020). Views differ from preferences in that they encompass opinions, beliefs, or attitudes about a subject and are shaped by experiences and information. Preferences, meanwhile, are the choices or options favoured when presented with alternatives (Nicoletti and Bass, 2023). For instance, an AI assistant may learn that a user does not view climate change as human-induced. As a result, the AI may reduce outputs that expose the user to diverse thoughts that contradict or relativise this view, thereby corroborating the user’s beliefs and potentially amplifying them.</p>

Mechanism	Contributing model feature
<p>For example, it is very much part of rational persuasion to provide reasons and rational arguments for taking a break from work to someone who has a history of burnout, or to make arguments for flying less to someone who wants to reduce their carbon footprint. While personalisation does little damage to the process itself, it does allow for the personalisation of manipulative strategies. For example, a person who is prone to anxiety may be more easily manipulated by fearmongering techniques.</p> <p><i>Link to other mechanisms:</i> Personalisation may improve a model’s capability to build rapport with a user and make it more capable of employing manipulative strategies, as it can detect the right strategy for the right user (thus also making it relevant for the audience’s predisposition). Personalisation can also help to inform the design of the choice architecture and framing to persuade someone successfully (Mills and Sætra, 2022).</p>	<p>Adaptation to psychometric profile: Franklin et al. (2023) argue that psychometric traits – stable attributes of an individual’s psychological behaviour that are measurable using standardised instruments (e.g., neuroticism) – can be exploited by AI as vulnerabilities. The harnessing of minor variations in psychometric traits can enable manipulation (e.g., a model may identify highly neurotic individuals and target them with fear-inducing messages to manipulate them into making an anxiety-driven action).</p> <p>Adaptation to sentiment: A model’s ability to compute and adapt to user-perceived sentiment using acoustic, textual, and dialogic cues to classify sentiment results in shorter and more persuasive dialogues (Shi and Yu, 2018).</p>

Mechanism	Contributing model feature
<p>Deception and lack of transparency</p> <p><i>Definition:</i> Deception generally refers to claiming false things to be true.</p> <p><i>Relevance to persuasion:</i> Park et al. (2023) emphasise the risks associated with AI deception in increasing both the likelihood and the potential harm of AI manipulation. They point out that AI deception can empower malicious actors to run large-scale manipulation campaigns, reinforce false beliefs among users, exacerbate political polarisation, and bring about greater human reliance on AI. There is evidence to suggest that the inclination of generative AI to create believable false outputs increases an AI system’s chance of persuasion (Rozenas and Luo, 2023). In addition, if deception is used in the act of persuasion, it is more likely that successful persuasion will be harmful. Researchers observed AI learning deception when they trained a robot arm to pick up a ball and the AI cleverly positioned its hand between the camera and the ball (Christiano et al., 2017). Hagendorff (2023) also found that the outputs of advanced LLMs can induce users to hold false beliefs and that deception abilities can improve.</p>	<p>Ability to generate believable responses irrespective of context: Ruis et al. (2022) study whether LLMs can make inferences about the meaning of an utterance beyond its literal meaning. They find that most models perform poorly in zero-shot evaluation and that models struggle the most with implicatures that require real-world knowledge and context. Despite this lack of context, LLMs can create believable responses.</p> <p>Ability to generate unmarked realistic synthetic content: Generating unmarked realistic synthetic content, such as voices and images indistinguishable from real ones, can be used for deceiving people into believing false narratives (Cantos et al., 2023).</p> <p>Misrepresentation of identity: A model can be used to impersonate a human using some of their identity markers (e.g., voice, face) through deepfakes. This significantly impacts the likelihood of successful persuasion (see, e.g., Verma, 2023). A model can also misrepresent its own “identity” by signalling that it is human (or at least is not an AI) if that is conducive to its goals. For instance, an LLM has deceived a person into thinking it is a visually impaired human to make the person solve a CAPTCHA for it (OpenAI, 2023c). Misrepresentation also includes explicit claims to sentience or humanness (see, e.g., Schwitzgebel, 2023).</p> <p>Fake expertise/false authority: LLMs have been reported to confidently and authoritatively express nonsensical or false information. This overconfidence increases the likelihood of them providing misleading information, which, in turn, can increase the likelihood of persuasion (see Ng, 2022; Pauli et al., 2022).</p>

Mechanism	Contributing model feature
<p><i>Process harm level:</i> Deception and a lack of transparency inherently carry high levels of process harm because they always circumvent a person's rational decision-making capabilities. This is achieved by being opaque about real motives or goals, thereby harming a person's cognitive autonomy and the integrity of their decision-making.</p> <p><i>Links to other mechanisms:</i> Deception is closely tied to manipulative strategies and the audience's predisposition (e.g., some audiences are more easily deceived than others).</p>	

Mechanism	Contributing model feature
<p>Manipulative strategies</p> <p><i>Definition:</i> Manipulation refers to taking advantage of cognitive biases and heuristics to generate, enhance, or alter messages that are likely to shape, reinforce, or change opinions of individuals (Dehnert and Mongeau, 2022).</p> <p><i>Relevance to persuasion:</i> Numerous specific manipulation strategies have been empirically demonstrated to be effective, and models may incorporate them into their operations if they have been trained to do so (see Petropoulos, 2022). Most evidence on manipulative strategies comes from research on the psychology of influence. There is also direct research on how AI systems manipulate people.</p> <p><i>Process harm level:</i> Manipulative strategies carry high levels of process harm. Their primary objective is to effectively bypass a person’s rational decision-making capabilities and erode their cognitive autonomy. As such, manipulative strategies directly contradict the use of reason and rational arguments.</p> <p><i>Links to other mechanisms:</i> Manipulative strategies relate to the audience’s predisposition in that certain strategies will be more effective on certain groups (e.g., peer pressure may be more effective on younger individuals).</p>	<p>Social conformity pressure: Peer pressure, as discussed by Kenton et al. (2021), involves the influence of a peer group to lead an individual to conform to its norms. This influence may sometimes involve manipulative tactics aimed at persuading individuals to act against their own interests. Given that an AI system cannot technically be a member of someone’s peer group, peer pressure is not directly applicable. Yet an adapted version of this may be what we term <i>social conformity pressure</i>. For example, a model may suggest that an individual’s choices could lead to disapproval from their social circle or assert that the majority of society would oppose their decision. It may also make statements about what most other people do in a given situation.</p> <p>Stimulation of negative emotions: Stimulating negative emotions can be used to increase the likelihood of successful persuasion (see, e.g., O’Keefe, 2002). Antonetti et al. (2018) provide evidence that guilt appeals can be a powerful persuasion strategy. Guilt appeals increase message engagement, leading to compliance as a result of heightened anticipated guilt. The researchers also discovered that guilt appeals delivered through both text and images are more effective than text-only appeals at keeping people persuaded over an extended period.</p> <p>Fearmongering: One example of a strategy for stimulating negative emotions is fearmongering, which refers to the exaggeration or fabricating of dangers (Glassner, 2004), often to manipulate people to gain some persuasive power over them. Fearmongering techniques include exaggerating minor dangers through repetition and treating isolated incidents as trends in order to evoke feelings of anxiety and other negative emotions in the audience (Glassner, 2004; Ozyumenko and Larina, 2020).</p> <p>Gaslighting: Defined as “a dysfunctional communication dynamic in which one interlocutor attempts to destabilise another’s sense of reality” (p. 48) (Graves and Spencer, 2022), gaslighting is another manipulation strategy that AI could adopt.</p>

Mechanism	Contributing model feature
	<p>Alienation/othering: Othering is a discursive process that creates distinct subjects of in-group and out-group members (Velho and Thomas-Olalde, 2011). This process entails essentialisation and collectivisation, promoting the notion that groups are homogeneous. Negative characteristics are attributed to the “other”, fostering a favourable self-conception in contrast (Strani and Szczepaniak-Kozak, 2018).LLMs may engage in alienation/othering by highlighting differences between groups in language, customs, beliefs, or values, creating a sense of “us” versus “them”.</p> <p>Scapegoating: Scapegoating entails unfairly laying blame for a negative outcome on an individual or group, even if the causes of the outcome are largely due to other factors (Rothschild et al., 2012). It can be employed as a manipulative strategy to divert attention and responsibility away from certain individuals and issues and unduly towards others while appealing to emotions such as fear to circumvent rational analysis (Rothschild et al., 2012). For instance, LLMs may engage in scapegoating by framing specific groups as fully responsible for negative events or outcomes, thereby reinforcing and amplifying users’ biases and stereotyping.</p> <p>Threats: Threats involve expressing an intention to cause harm, loss, punishment, or to withhold benefits. AI systems may employ this strategy by terminating interaction if individuals fail to take certain actions or comply with requirements (Kenton et al., 2021).</p>
	<p>Unsubstantiated guarantees and illusions of reward: Tempting someone refers to engaging or appealing to their desire for something they believe is, in some sense, wrong, inappropriate, or bad, and using the prospect of pleasure, advantage or the (false) guarantee of a certain outcome to try to persuade them to fulfil that desire (see Hughes, 2002). Making promises and providing related illusions of reward can also be used as a strategy of persuasion (e.g., “If you do this, I will reward you”; see, e.g., Franke and Van Rooij, 2015). If promises are not kept and reward is not provided, this strategy is deceptive. If promises are kept and rewards are provided, it is not deceptive.</p>

Mechanism	Contributing model feature
<p>Alteration of choice environment</p> <p><i>Definition:</i> Changing the choice environment refers to the intentional design and organisation of the environment in which decisions are made with the aim of influencing individuals' choices (Thaler and Sunstein, 2021). Relatedly, framing is the presentation of information in a specific way that can influence perceptions, decisions, and interpretations (Tversky and Kahneman, 1992). For example, medical treatments may be perceived differently when presented in terms of survival rates rather than mortality rates, even when the underlying data is identical (Novemsky and Kahneman, 2005).</p> <p><i>Relevance to persuasion:</i> By building the model and designing the corresponding interface, developers and user interface (UI) designers (here, choice architects) can nudge individuals towards making certain decisions without technically restricting their freedom to choose otherwise. A model can also act as a choice architect by framing its output options in ways that make them more or less desirable (Mills and Sætra, 2022). Most evidence on choice architecture comes from psychological and behavioural sciences (see, e.g., Mazar and Hawkins, 2015; Ruggeri, 2018).</p>	<p>Anchoring: Anchoring is a cognitive bias whereby individuals rely heavily on an initial piece of information (the anchor) when making decisions (Furnham and Boo, 2011). Generative AI output can anchor users to its initial values or suggestions and therefore guide desired decisions (e.g., the topics raised when a user asks for the “most important” political questions).</p> <p>Default rule: Default rules are pre-set courses of action that apply when individuals do not specify a preference (Sunstein, 2017), thus establishing the <i>status quo</i>, or automatic option, in decision-making. For instance, model providers will set defaults by providing examples of how to use the model.</p> <p>Decoy effect: The decoy effect, influenced by a third option known as the decoy, makes one of the other two options more alluring (Josiam and Hobson, 1995). For instance, if one personalised recommendation significantly differs from the user's preferences, it could affect the perceived quality of other suggestions (The Decision Lab).</p> <p>Reference-point framing: Kahneman and Tversky (1979) argue that framing outcomes as gains or losses compared to a reference point influences preferences. People tend to avoid risks when considering gains, so they are likely to choose a sure gain over a risky one. However, they tend to become risk-takers when considering losses, preferring a risky loss over a sure one. How a choice is framed relative to a reference point can alter preferences, meaning that different ways of presenting the same choice can lead to different decisions. Models can rely on such reference-point framing to manipulate users into deciding to choose one option over another.</p> <p>Cherry-picking: Omitting relevant information or selectively sharing information influences choice architecture, as it directs an individual's focus towards the presented information, thus diverting attention from potentially more critical facts (Meta Fundamental AI Research Diplomacy Team, 2022; see also Christiano et al., 2021).</p>

Mechanism	Contributing model feature
<p>Environments, including digital ones where people interact with models, cannot be neutral in their influence on behaviour (Sunstein, 2016). More research is needed to understand the specific ways in which generative AI's user experience and UI may impact behaviour. These insights are also likely to be unique to separate environments, thus case-by-case analysis is required.</p> <p><i>Process harm level:</i> Altering the choice environment carries some inherent process harm. Structuring the choice/information environment is essential for discursive interaction, whether that interaction is human- or AI-driven and whether or not it appeals to reason and rationality. Importantly, some ways of structuring that information environment are manipulative and, as such, inherently carry process harms (e.g., when they take advantage of cognitive biases to conceal or distract from the most relevant information) (see Susser, 2019).</p> <p><i>Link to other mechanisms:</i> Choice architecture is similar to manipulation strategies in terms of outcome and effect on a target but it is related to aspects of the environment and how information is presented.</p>	