Thinking Outside the Technical Standardisation Box: The Role of Standards Under the Draft EU Artificial Intelligence Act

Tristan Goodman*

ABSTRACT

A lack of standardisation is a significant barrier to wider adoption of artificial intelligence (AI) in society. Without the establishment of commonly accepted standards for the development and deployment of AI systems, technical barriers will remain. Unexpected outcomes, in turn, are likely to generate distrust. This will result in significant socio-economic opportunity costs. In response, recent years have seen a proliferation of AI standardisation initiatives across multiple jurisdictions. As part of these efforts, the EU has taken a relatively unique approach by relying on private European standardisation organisations (ESOs) to develop technical standards which are intended to operationalise the legal requirements for certain AI systems contained in its proposed Artificial Intelligence Act. Despite this reliance, this article demonstrates how operationalising these requirements requires answering a series of hard normative questions that ESOs are unlikely to answer. While much scholarly and political attention has focused on reforming and modifying European technical standardisation in anticipation of these shortcomings, this article suggests that we should think outside this 'technical standardisation box'. Another less explored standardisation process should be considered instead: the development of regulatory conventions through regulatory sandboxing.

*LLM (London School of Economics and Political Science) '23. LLB (University of Bristol) '17. Any mistakes are my own.

INTRODUCTION

Shortly after the global outbreak of COVID-19, researchers scrambled to develop artificial intelligence (AI) tools to help medical practitioners diagnose the disease and determine its risk to patients.[1] Although these considerable efforts highlighted the potential benefits of AI in socially important domains such as healthcare, the results were disappointing. Several academic studies concluded that, in practice, these tools were unhelpful and, in some cases, potentially harmful.[2] Many of the identified issues related to the poor quality of the data that researchers used to develop their models, as well as deficiencies in methodology and reporting. For example, many researchers had used data from paediatric patients without COVID-19 to train their models to detect the disease in adults, resulting in AI tools that learned how to detect children rather than COVID-19.[3] Central to these issues was a lack of reporting on the demographic properties of training data. Indeed, this generated wider concerns among researchers that AI tools were being developed without due consideration of sampling biases, such as the insufficient representation of minority groups that have been

[1] Will Douglas Heaven, 'Hundreds of AI tools have been built to catch covid. None of them helped.' (*MIT Technology Review*, 30 July 2021) <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/> accessed 18 September 2023.

[2] Laure Wynants and others, 'Prediction models for diagnosis and prognosis of covid-19: systemic review and critical appraisal' (2020) 369 BMJ <https://www.bmj.com/content/369/bmj.m1328> accessed 18 September 2023; Michael Roberts and others, 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans' (2021) 3 Nature Machine Intelligence 199; Alan Turing Institute, 'Data science and AI in the age of COVID-19' (June 2021). <https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf> accessed 18 September 2023.

[3] Roberts and others (n 2) 211.

disproportionately affected by COVID-19.[4] A key contributing factor to these issues was the lack of adequate data standardisation: '[d]ifferent data standards and codification of metadata, and lack of dataset documentation, meant that data were difficult to find, link and assess in terms of missingness and biases, limiting the scope of and confidence in analyses'.[5]

Examples like this highlight how the lack of standardisation — that is, an agreed, repeatable way of doing something[6] — remains a key barrier to the broader implementation of AI in society.[7] Without commonly accepted standards on how to develop and use AI systems, technical barriers will constrain innovation whilst unexpected outcomes will generate distrust, resulting in significant socio-economic opportunity costs. The importance of standards in the development and use of AI is reflected in the proliferation of public-private AI standardisation initiatives in recent years at both the national and international levels.[8]

As part of these standardisation efforts, the EU is taking a relatively unique approach by developing standards to aid the implementation of key provisions in its proposed Artificial Intelligence Act.[9] Forming part of the EU's

---

[4] Alan Turing Institute (n 2) 14.
[5] Alan Turing Institute (n 2) 12.
[6] British Standards Institution, 'Information about standards' <https://www.bsigroup.com/en-gb/standards/Information-about-standards/> accessed 18 September 2023.
[7] For the same example, see: Johann Laux, Sandra Wachter, and Brent Mittelstadt, 'Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act' (Oxford Internet Institute, 2023), 3 <https://ssrn.com/abstract=4365079> accessed 18 September 2023.
[8] There are currently over 300 AI-related standards which prominent standards development organisations have either published or are currently developing. For a searchable database, see: AI Standards Hub, 'Standards Database' <https://aistandardshub.org/ai-standards-search/> accessed 18 September 2023.
[9] Commission Proposal 2021/0106 of 21 April 2021 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 [2021]

product safety regime, the Draft AI Act is partly modelled on the 'New Legislative Framework' (NLF). This allows the European Commission to request European standardisation organisations (ESOs) to develop standards, also known as 'harmonised standards', to translate legal requirements into 'objectively verifiable criteria' against which AI systems can be certified as conforming with NLF legislation.[10] In the words of the Commission, harmonised standards are intended to provide the 'precise technical solutions' for regulatory compliance.[11]

Whilst this partnership between regulation and standardisation has worked effectively in the context of other products, such as toys and explosives, it is likely to become strained in the context of AI systems.[12] While NLF legislation typically regulates products that pose risks to human life or physical injury, AI systems pose additional risks to fundamental rights and related public interests.

COM/2021/206 (Draft AI Act). In this article, references to the Draft AI Act are references to the text proposed by the Commission rather than references to the texts which the Council of the European Union and the European Parliament have subsequently proposed unless otherwise stated. For the text proposed by the Council, see: Council Proposal 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach' (25 November 2022) <https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf> accessed 18 September 2023. For the text proposed by the Parliament, see: 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)) [2023] P9_TA(2023)0236.

[10] CEN and CENELEC, 'Drafting Harmonized Standards in support of the Artificial Intelligence Act (AIA)' (21 March 2022) <https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN-CENELEC_Topics/Artificial%20Intelligence/jtc-21-harmonized-standards-webinar_for-website.pdf> accessed 18 September 2023.

[11] Commission Proposal (n 9) 13; Commission Notice 2022/3637/EC of 29 June 2022, The 'Blue Guide' on the implementation of EU product rules 2022 [2022] OJ C 272/1 (Blue Guide), para 4.1.1.

[12] For a compilation of current NLF legislation, see: European Commission, 'New Legislative Framework' <https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en> accessed 18 September 2023.

These are typically less readily quantifiable and more open to interpretation.[13] To evoke the opening example, it is relatively uncontroversial to claim that AI tools should provide reliable outputs for patients irrespective of their demographical status. However, determining what constitutes a 'sufficiently representative' dataset and what level of privacy intrusion would be acceptable to achieve this outcome is much more contestable.[14] These decisions involve trade-offs between competing interests and/or endorsements of specific interpretations of contested concepts — what Laux and others call 'hard normative questions'[15] — that cannot be reduced to 'objectively verifiable criteria' or 'precise technical solutions'. Instead, these rely on the judgement of local decision-makers.

At a time when EU legislators are negotiating agreement on the final version of the Draft AI Act,[16] the apparent disconnect between the nature of harmonised standards and the legal requirements they are intended to operationalise raises important questions for regulators in the EU and beyond. What are the kinds of hard normative questions that those developing and using AI systems face? To what extent have EU legislators provided guidance? Can, and should, ESOs play a role in answering these questions? If not, what is the alternative? Without thoughtful answers to these questions, policymakers cannot

---

[13] Christine Galvagna, 'Discussion paper: Inclusive AI governance — Civil society participation in standards development' (*Ada Lovelace Institute*, 30 March 2023), 17-18 <https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/> accessed 18 September 2023.

[14] As discussed in Part I, section B(ii).

[15] Laux, Wachter, and Mittelstadt (n 7). In this article, the phrase 'hard normative questions' is adopted as short-hand for decisions involving trade-offs between competing interests and/or specific interpretations of contested concepts, in the same way as Laux, Wachter, and Mittselstadt.

[16] The Parliament, the Council, and the Commission are currently engaged in inter-institutional negotiations — known as 'trilogues' — to reconcile their respective proposed texts. The final version of the AI Act is expected around the end of 2023.

properly assess what role standards should play under the Draft AI Act and other emerging regulatory frameworks for AI.[17]

This article attempts to provide some answers to these questions in Parts I, II, and III. Part I begins by illustrating how essential legal requirements in the Draft AI Act raise hard normative questions which are left largely unanswered by EU legislators. Part II then considers the approach that ESOs are likely to take when developing harmonised standards. It argues that, based on this approach, resulting standards are unlikely to provide objectively verifiable criteria for AI providers to satisfy or precise technical solutions to use when navigating hard normative questions. Much scholarly and political attention has focused on reforming or modifying European technical standardisation in anticipation of these shortcomings.[18] Part III, however, concludes by suggesting that we should think outside this 'technical standardisation box' and consider another, less explored standardisation process: the development of regulatory conventions through regulatory sandboxing.

---

[17] Expected to be the world's first comprehensive AI regulatory regime, the AI Act is likely to influence regulation adopted in other jurisdictions — known as the 'Brussels Effect'. For example, see: Charlotte Siegmann and Markus Anderljung, 'The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market' (August 2022) <https://doi.org/10.48550/arXiv.2208.12645> accessed 18 September 2023.

[18] See for example: Mark McFadden and others 'Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation' (*Oxford Information Labs*, December 2021) <https://oxil.uk/publications/2021-12-02-oxford-internet-institute-oxil-harmonising-ai/Harmonising-AI-OXIL.pdf> accessed 18 September 2023; Galvagna (n 12); Laux, Wachter and Mittelstadt (n 7); Hans-W Micklitz, 'The Role of Standards in Future EU Digital Policy Legislation' (July 2023) <https://www.beuc.eu/sites/default/files/publications/BEUC-X-2023-096_The_Role_of_Standards_in_Future_EU_Digital_Policy_Legislation.pdf> accessed 18 September 2023.

I. THE POLITICS OF AI REGULATION

Although the field of AI was established in the 1950s,[19] its real-world applications have only accelerated in recent years with the advent of novel machine learning (ML) techniques, such as deep learning (DL), which have rapidly advanced with increases in the availability of data, computing power, and investment.[20] Like other rapidly developing technologies, AI can improve aspects of people's lives in ways which were not previously possible, whilst also posing risks in ways which are not always foreseeable.[21] This is particularly true in the case of advanced AI systems: ML techniques, which enable the augmentation and even the replacement of human expertise in a variety of socially important domains, are also what can make those systems inexplicable and unpredictable. As further discussed in Section B of this Part, these characteristics of AI systems raise a series of hard normative questions.

As the role of AI systems in people's lives has grown, so has concern about how these hard normative questions are answered. The public and regulators are increasingly aware that private organisations are making value judgements which can have significant societal impacts. At the same time, those

---

[19] For a brief overview of the history of AI and the distinct intellectual roots of ML, see: Harry Law, 'An introduction to AI history' (*Learning From Examples*, 7 August 2023) <https://learningfromexamples.substack.com/p/an-introduction-to-ai-history> accessed 18 September 2023.

[20] Miles Brundage and others, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (*arXiv*, February 2018), 12 <https://arxiv.org/abs/1802.07228> accessed 18 September 2023.

[21] David Leslie, 'Understanding artificial intelligence ethics and safety: A guide to AI ethics, including responsible design and implementation of AI systems in the public sector' (*Alan Turing Institute*, 11 June 2019), 3 <https://doi.org/10.5281/zenodo.3240529> accessed 18 September 2023.

developing and deploying AI systems are increasingly at risk of negative economic impacts when their systems produce outcomes which are misaligned with the expectations of the public, regulators, and even developers.[22] In response, the AI industry and governmental bodies from around the world are creating and implementing standardisation initiatives which are intended to provide those developing and deploying AI systems with the tools and processes they need to meet societal and regulatory expectations, as well as their own.[23]

This Part critically assesses the approach that the EU has taken as part of these global AI standardisation efforts. Doing so requires a brief explanation of the regulatory theory which underpins the EU approach,[24] since it will be argued that the divergence of the Draft AI Act from its theoretical foundations is a root cause of many of the challenges which technical standardisation is likely to face in reality.

A. Standardising regulation: the theory

For many decades, EU legislators have refrained from detailing technical requirements for a variety of products in legislation, limiting themselves to setting

[22] Juan Aristo Baquero and others, 'Derisking AI by design: How to build risk management into AI development' (*QuantumBlack AI by McKinsey*, 13 August 2020) <https://www.mckinsey.com/capabilities/quantumblack/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development> accessed 18 September 2023.

[23] Hadrien Pouget, 'What will the role of standards be in AI governance? Why standards are at the centre of AI regulation conversations and the challenges they raise' (*Ada Lovelace Institute*, 5 April 2023) <https://www.adalovelaceinstitute.org/blog/role-of-standards-in-ai-governance/> accessed 18 September 2023.

[24] For a seminal account of the law on standards in the EU, see: Harm Schepel, *The Constitution of Private Governance* (1st edn, Hart Publishing 2005). For a recent revival, see: Olia Kanevskaia, *The Law and Practice of Global ICT Standardization* (Cambridge University Press 2023).

out essential health and safety requirements for those products to have access to the internal market. With the aim of operationalising these essential requirements, the Commission can request recognised private standardisation organisations — the ESOs — to draft technical specifications in the form of 'European standards'. These standards are developed in accordance with the procedure laid down by the Standardisation Regulation,[25] and, assuming they are approved by the Commission, eventually cited in the Official Journal of the EU as 'harmonised standards'.[26] Although manufacturers can use alternative means to comply with applicable essential requirements, products that comply with harmonised standards cited in the Official Journal benefit from a presumption of conformity.[27] In practice, this means that compliance with harmonised standards is the easiest and safest route to ensure that a product meets EU regulatory requirements.[28]

This style of regulation, originally formulated as the 'New Approach',[29] was prompted in the 1980s by a perceived need to ensure that EU regulation

[25] Council Regulation (EC) 2015/2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council [2012] OJ L316/12 (Standardisation Regulation).

[26] In practice, the Commission frequently declines to cite potential harmonised standards in the Official Journal, usually on the advice of private consultants — known as harmonised standards or 'HAS' consultants — who determine whether harmonised standards satisfy EU law requirements. See: Commission Report 2022/30 from the Commission to the European Parliament and the Council on the implementation of the Regulation (EU) No 1025/2012 from 2016 to 2020 [2022] COM (2022) 30 final, para. 2.7.1.

[27] Commission, 'Blue Guide' (n 11) para. 4.1.2.

[28] Rob van Gestel and Hans-W Micklitz, 'European Integration Through Standardization: How Judicial Review is Breaking Down the Club House of Private Standardization Bodies' (2013) 50 Common Market Law Review 145, 157.

[29] Despite its name, there is nothing 'new' about the NLF, which is based on EU case law from the 1970s. See: Commission, 'Blue Guide' (n 11) 5-8.

could respond to the pace of innovation in many industries.[30] Private regulation, including standardisation, is generally more flexible and responsive than the EU legislative process.[31] It also benefits from a high level of technical expertise, since industry representatives have up-to-date knowledge and are generally better equipped to deal with complex technical issues than legislators.[32] As the pace of innovation continues to increase, so too does the demand for technical standards as a regulatory tool.[33]

However, delegating regulatory authority to ESOs inevitably raises constitutional challenges.[34] Since ESOs are made up of industry representatives and are governed by private law, they are neither representative of, nor accountable to, the public in the same way that legislators are. In response to this, recital 5 of the Standardisation Regulation and official guidance published by the Commission emphasise that requirements set out in NLF legislation and related standardisation requests should be precise. Indeed, sufficient precision is required to avoid misinterpretation on the part of ESOs to ensure that political choices are

[30] Commission, 'Blue Guide' (n 11) para 1.1.1. Another, more historical driver was that the delegation of standard-setting to private rule makers paid lip service to both the subsidiary principle and the demands of business to cut red tape, at a time when the first wave of de-regulation had recently swept across Europe. See: van Gestel and Micklitz (n 28) 156.

[31] Herwig C.H. Hofmann, Gerard C. Rowe, and Alexander H. Türk, *Administrative Law and Policy of the European Union* (Oxford University Press 2011), 12.

[32] Mislav Mataija, *Private Regulation and the Internal Market* (Oxford University Press 2016), 12.

[33] There are now more than 3,600 harmonised European standards. In response to the increasing demand for standards, the Commission is now reviewing how the European standardisation process can be further streamlined and become a more responsive part of its latest 'Standardisation Strategy', published in February 2022. See: Commission, 'An EU strategy on Standardisation — Setting global standards in support of a resilient, green and digital EU single market' COM (2022) 31 final.

[34] See generally: Mariolina Eliantonio and Caroline Cauffman, *The Legitimacy of Standardisation as a Regulatory Technique* (Edward Elgar Publishing 2020).

not left to them.[35] According to the Commission, such choices include determining the maximum level of exposure to a hazard.[36] For example, the NLF legislation regulating toys specifies the maximum acceptable level of exposure which users of toys can have to specific hazardous chemicals transferring from a toy's materials.[37] The 'precise technical solutions' for ensuring that toys do not exceed these maximum exposure levels are set out in harmonised standard EN 71-3:2019+A1:2021.[38]

However, in reality, this theoretical division of labour is likely to become strained in the context of AI systems, where value judgements and technical decision-making are often inextricably linked. The remainder of this Part will demonstrate how the EU regulatory approach to AI systems leaves these and other related political choices for others to make, even though the Commission specifically uses the maximum level of hazard exposure as an example of a political choice which must remain with legislators.

B. Standardising AI regulation: the reality

Arguably, two of the most important actors in the Draft AI Act are the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC).[39] These constitute two of the

[35] Commission Staff Working Document Ref. Ares(2015)4888382 of 6 November 2015 Vademecum on European Standardisation in support of Union legislation and policies [2-15] SWD(2015) 205 final, Part 1/3 (Vademecum on European Standardisation), 9; Commission, 'Blue Guide' (n 11) para 4.1.1.
[36] Commission, 'Vademecum on European Standardisation' (n 35) 8-9.
[37] European Parliament and Council Directive 2009/48/EC of June 2009 on the safety of toys [2009] OJ L170/1, Annex II, Part III, point 13.
[38] CEN, *Safety of toys – Migration of certain elements (EN 71-3:2019+A1:2021)* (British Standards Institution 2021).
[39] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 Computer Law Review International 97, 104.

three ESOs that have their standards recognised as 'European standards'.[40] Although neither is mentioned in the text, both will have the responsibility of developing harmonised European standards for the essential requirements for 'high-risk' AI systems contained in Title III, Chapter 2 of the Draft AI Act.[41] This was subsequently confirmed by the Commission in a draft standardisation request issued to CEN and CENELEC on 5 December 2022.[42]

By analysing the provisions of the essential requirements below, it will be shown that the regulatory power afforded to CEN and CENELEC is insufficiently curtailed by ambiguous and high-level drafting, thereby leaving many of the hard normative questions raised by AI systems unanswered.[43]

---

[40] Annex I of the Standardisation Regulation provides an exhaustive list of recognised ESOs: CEN (European Committee for Standardisation), CENELEC (European Committee for Electrotechnical Standardisation) and ETSI (European Telecommunications Standards Institute).

[41] 'High-risk' AI systems are defined in Section 5.2.3 of the Draft AI Act. As the Explanatory Memorandum of the Draft AI Act summarises, they comprise AI systems which are deemed to 'create a high risk to the health and safety or fundamental rights' of individuals'. Determining whether an AI system is 'high-risk', and the difficulties of making such determinations, is not considered in this article.

[42] Commission, 'Draft standardisation Request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence' (5 December 2022) (Draft Standardisation Request) <https://ec.europa.eu/docsroom/documents/52376> accessed 18 September 2023. Since the Draft AI Act is not yet in force, let alone agreed, the Draft Standardisation Request represents an initial stage in the standardisation process, whereby CEN and CENELEC develop 'technical specifications' which are intended to form the basis of future harmonised standards. See: Draft Standardisation Request (n 42) recital 15; CEN and CENELEC (n 10).

[43] The drafting of the requirements contained in Annex II of the Draft Standardisation Request is not considered here, since, in the words of one commentator, those requirements 'are an incomplete version of the more detailed and more specific legal requirements' in Title III, Chapter 2 of the Draft AI Act: Micklitz (n 18) 137. This is a deficiency in itself, but in terms of drafting specifically, it can be assumed that deficiencies in the Draft AI Act are equally present, if not more deficient, in the Draft Standardisation Request.

Whether ESOs can or should be able to provide those answers will then be considered in Part II.

(i) Risk management systems

According to the Explanatory Memorandum of the Draft AI Act, AI systems may pose a risk to a wide range of fundamental rights under the Charter of Fundamental Rights of the European Union.[44] This includes rights to human dignity, respect for private life, protection of personal data, non-discrimination, equality between women and men, freedom of expression and assembly, an effective remedy and a fair trial, and so on.[45] To help mitigate these risks, Article 9(2)(a) of the Draft AI Act provides that risk management systems must be established for high-risk AI systems which identify, estimate, and evaluate known and 'reasonably foreseeable' risks to the fundamental rights of individuals when the system is 'used in accordance with the intended purpose and under conditions of reasonably foreseeable use'.

One of the clearest challenges to technical standardisation presented by this essential requirement is the need to determine 'reasonably foreseeable' use cases of an AI system. The foreseeability of use cases for AI systems are incomparable with that of other products regulated by NLF legislation, such as toys and fireworks, where risks are largely a matter of statistics and studying household accidents.[46] As a general-purpose technology which is capable of continual adaptation, potential uses of AI systems are effectively endless and are often only limited by human imagination. The rapidly developing, and sometimes

[44] Charter of Fundamental Rights of the European Union [2000] OJ C364/01 (CFEU).
[45] CFEU (n 44); Commission (n 11) 11.
[46] Micklitz (n 18) 94.

unexpected, applications of so-called 'foundation models', such as ChatGPT, provide a recent example.[47]

However, even if it were possible to standardise foreseeable use cases, identifying what 'reasonably foreseeable' risks are in light of those cases implies hard normative questions. For example, do these risks only affect individuals, or do they apply to groups as well? While Article 9(8) of the Draft AI Act sheds some light on this by providing that 'specific consideration' is given to children, it does not elaborate on whether any other groups should be considered. It is plausible that other groups would be deemed to need protection under the CFEU due to risks of exclusion or abuses of power, such as people with disabilities, the elderly, and workers.[48] However, the Draft AI Act does not clarify this. Even if it did, identifying members of these groups is not, in general, objectively verifiable, since the make-up of disadvantaged groups is typically decided on a case-by-case basis in EU law.[49]

Hard normative questions also arise in the evaluation of identified risks to fundamental rights. Unlike risks to human life or physical injuries, risks to fundamental rights are multi-dimensional, thereby making it unfeasible to develop

[47] Markus Anderljung and others, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (2023) Cornell University, 15 <https://arxiv.org/abs/2307.03718> accessed 18 September 2023.

[48] As suggested by the independent High-Level Expert Group on Artificial Intelligence in its 'Ethics Guidelines for Trustworthy AI', from which the essential requirements for high-risk AI systems in the Draft AI Act derive (High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019), 2 <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 18 September 2023.

[49] Christopher McCrudden and Sacha Prechal, 'The Concepts of Equality and Non-Discrimination in Europe: A Practical Approach' (2009) Oxford Legal Studies Research Paper No. 4/2011, 35-36.

a single, generalisable metric to measure them.[50] This would be the case if, for example, those developing the AI tools for the diagnosis of COVID-19 wished to evaluate the level of risk their models posed to patients' right not to be discriminated against based on protected characteristics, such as race or age. This could be done by measuring the distribution of outcomes across different demographics from the predictions of the AI tool according to several different metrics.[51] For example, developers could measure the rate at which the system incorrectly identifies a patient as having COVID-19 across different groups (known as 'false positive rate equality'). Alternatively, the accuracy of the AI system's predictions — for example, whether the rate at which the system incorrectly predicts a patient to have COVID-19 is the same for black and white patients (known as 'predictive parity') — could be measured. While both metrics are reasonable, they cannot be satisfied simultaneously in plausible scenarios,[52] such as where the prevalence of COVID-19 differs across demographics. Additionally, the AI system cannot make perfect predictions. In such a scenario, an AI system that provides predictions which are equally accurate between different groups will inevitably produce more false positives for the group with a higher prevalence of COVID-19, since it will be disproportionately affected by the overall imperfect predictive capability of the AI system. Violation of either false positive rate equality or predictive parity is, therefore, *potential* rather than conclusive evidence of discrimination.[53] Determining whether that evidence

[50] Galvagna (n 13) 39.
[51] Twenty distinct metrics are considered in: Sahil Verma and Julia Rubin, 'Fairness Definitions Explained' (May 2018) ACME/IEEE Workshop on Software Fairness <https://doi.org/10.1145/3194770.3194776> accessed 18 September 2023.
[52] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' (2016) <https://arxiv.org/abs/1609.05807> accessed 18 September 2023.
[53] Deborah Hellman, 'Measuring Algorithmic Fairness' (2020) 106 Virginia Law Review 811, 836-837; Will Fleisher, 'Algorithmic Fairness Criteria as Evidence' [2023] Social Science Research Network 1, 5-8

constitutes an unacceptable risk is context-dependent; this will thereby rely on the judgement of local decision-makers.

Even if there were clear normative thresholds for fundamental rights risks, such as non-discrimination, mitigating them requires further subjective decision-making. For example, if an AI diagnostic tool produces discriminatory outcomes because it is trained on insufficiently representative data, then the technical solution could be to ensure that future training data is more representative for those groups which are currently disadvantaged by the AI tool. But what if a lack of data stems from more systemic inequalities, such as a lack of access or inability to use mobile devices? Addressing 'data privilege issues' like these would thereby likely require a broader policy response, rather than a technical solution alone.[54] While Article 9 of the Draft AI Act does not exclude different types of mitigation measures, it fails to provide clear guidance. Article 9(2)(d) requires that 'suitable risk management measures' are adopted, whereby any residual risk of high-risk AI systems is 'judged acceptable'. This type of normative inquiry, where competing interests involving fundamental rights and other public interests must be balanced, is typically the domain of the legislature or constitutional courts, rather than private organisations or individuals, such as ESOs and AI developers.

(ii) Data and data governance

Another hard normative question raised by the example above concerns the level of privacy invasion that is acceptable in order to collect additional data to ensure that training datasets are sufficiently representative. This is a particularly complex question, since the requisite data will often concern sensitive data relating

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3974963> accessed 18 September 2023.

[54] Alan Turing Institute (n 2) 15.

to protected characteristics, such as race or sex, which are given heightened protection under EU data protection law.[55] While Article 10(5) of the Draft AI Act suggests that 'special categories of data' may be processed for the detection and mitigation of biased training data where 'privacy-preserving' technical measures are in place, it does not elaborate on what those measures should be. Should these measures comply with those already provided in EU data protection legislation, such as pseudonymisation? What about the use of synthetic data to fill gaps in datasets? While amendments made by the European Parliament to the Draft AI Act elaborate on several minimum data governance requirements that derive from EU data protection law, the list provided is non-exhaustive.[56] For now, the Draft AI Act leaves many questions relating to the management of the trade-off between more representative data and invasions of privacy unanswered.

(iii) Accuracy and robustness

Even if AI systems are developed using sufficiently representative and relevant data, this does not guarantee that their predictions will be reliable. This is particularly true in the case of DL systems which use large amounts of data to subsequently recognise and classify related but previously unobserved data, using deep neural networks which are largely inscrutable and can learn from data in unintuitive ways.[57] Therefore, it is necessary that additional measures are taken to ensure that AI systems perform consistently throughout their lifecycle, thereby

[55] For example, see Article 9 of Council Regulation (EC) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.
[56] Amendments adopted by the European Parliament (n 9) Article 10(5).
[57] A famous example is a neural network system which learnt to distinguish between huskies and wolves based on whether there was a snowy background. See: Wojciech, Samek, and others, 'Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications' (2021) 109 Proceedings of the IEEE 247, 265 <https://arxiv.org/abs/2003.07631> accessed 18 September 2023.

reducing the risk of erroneous or biased outputs. On this basis, Article 15(1) of the Draft AI Act requires high-risk AI systems to achieve an 'appropriate level' of 'accuracy' and 'robustness' by implementing 'state-of-the-art measures'. Annex II of the Draft Standardisation Request clarifies that, in this context, 'state-of-the-art' should be understood as 'a developed stage of technical capability' which is 'accepted as good practice in technology'.

An immediate issue with this requirement is that it is unclear whether there are, in fact, any sufficiently mature and commonly accepted metrics for measuring accuracy or robustness.[58] Returning to neural networks, it is difficult to ensure that they function reliably because 'it is not always clear what should be tested for and observers are continually surprised by ways in which these systems fail'.[59] For example, not long after OpenAI released ChatGPT-3, a neural network chatbot, users found ways to manipulate it to produce potentially dangerous content. This was notwithstanding the fact that it had been developed and tested by leading researchers to avoid such outputs.[60] This problem is also likely to persist in the near future. As OpenAI note on their website, 'Despite its capabilities, GPT-4 […] [is still] not fully reliable' and 'there still exist "jailbreaks" to generate content which violates our usage guidelines'.[61]

However, even if it could be said that there is an accepted state-of-the-art, at least for some AI systems, applying metrics for accuracy and robustness

[58] Michael Veale, 'Value-Laden Areas for Standardisation in the AI Act' (5 September 2022) <https://michae.lv/value-laden-areas-in-the-ai-act/> accessed 18 September 2023.
[59] Hadrien Pouget, 'The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist' (*Lawfare*, 12 January 2023) <https://www.lawfaremedia.org/article/eus-ai-act-barreling-toward-ai-standards-do-not-exist> accessed 18 September 2023.
[60] ibid.
[61] OpenAI, 'GPT-4' (14 March 2023) <https://openai.com/research/gpt-4> accessed 18 September 2023.

still entails difficult trade-offs. For example, the accuracy of an AI system — understood in this context as the ability of a system to perform the task for which it has been designed[62] — can be measured by its specificity (its true negative rate) or by its sensitivity (its true positive rate).[63] Improving the specificity of an AI system can lead to a reduction in its sensitivity and vice versa.[64] Deciding which metric should be used to determine an acceptable level of accuracy is, therefore, context-dependent; the same can be said regarding the ultimate judgement on what level of accuracy is acceptable. For example, sensitivity would be more important where an AI system is used to detect COVID-19 in patients, since the potentially fatal consequences of false negatives are more important to minimise than false positives. Similar judgement calls are required in the context of robustness. For example, a computer vision AI system used for autonomous vehicles will need to perform reliably in new and adverse conditions, since the harms resulting from malfunction are potentially fatal.[65]

(iv) Transparency and human oversight

The difficulty of ensuring that certain AI systems perform reliably means that those using or monitoring such systems should be able to verify and control

[62] Rather than its narrower definition of statistical accuracy, which is one of several metrics for evaluating the performance of an AI system, as clarified in: Draft Standardisation Request (n 42) 4.

[63] The 'true negative rate' or 'specificity' of an AI system refers to the percentage of all cases which the system correctly identifies the absence of the property being predicted; inversely, the 'true positive rate' or 'sensitivity' refers to the percentage of all cases which the system correctly identifies the presence of the property being predicted.

[64] Thomas F Monaghan and others, 'Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value' (2021) 57 Medicina 503.

[65] Christian Berghoff and others, 'Robustness Testing of AI Systems: A Case Study for Traffic Sign Recognition' in Ilias Maglogiannis and others (eds) *Artificial Intelligence Applications and Innovations: IFIP Advances in Information and Communication Technology* (vol 627, Springer Cham 2021).

outputs in order to minimise the risks of unexpected system functioning. This is, in part, what drives the transparency and human oversight requirements provided in Articles 13 and 14 of the Draft AI Act.[66] Article 13(1) provides that high-risk AI systems must be developed such that their operation is sufficiently transparent to interpret the system's output and to use it. In simple terms, an AI system is interpretable if a human can readily understand the functional behaviour of the system without requiring additional tools.[67] Since interpretability is defined by reference to human cognition, for which there is no objective measure, there can be reasonable disagreement about whether or not an AI system is interpretable. That said, the simpler the AI system, the more interpretable it is likely to be.

However, many AI systems are too complex to be interpretable, even for experts.[68] This is known as the 'black box problem', which illustrates the impossibility of explaining the decision-making process of the AI system without

[66] In relation to transparency, see: Commission Proposal (n 9) s 3.5: 'In case infringements of fundamental rights still happen, effective redress for affected persons will be made possible by ensuring transparency'. For human oversight, Article 14(2) of the Draft AI Act provides: 'Human oversight shall aim at preventing or minimising the risks to […] fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of the other [essential] requirements.'

[67] Brent Mittelstadt, 'Interpretability and Transparency in Artificial Intelligence' in Carissa Véliz (ed), *The Oxford Handbook of Digital Ethics* (online edn, Oxford Academic 2021) <https://doi.org/10.1093/oxfordhb/9780198857815.013.20> accessed 18 September 2023.

[68] As a recent overview of the field of interpretable ML concluded, '[i]t remains unclear when – or even if – we will be able to deploy truly interpretable deep learning systems'. See: Tim GJ Rudner and Helen Toner, 'Key Concepts in AI Safety: Interpretability in Machine Learning' (*Center for Security and Emerging Technology*, March 2021), 6 <https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning/> accessed 18 September 2023.

additional explainability tools.[69] Therefore, will such systems fail to meet this essential requirement and not have access to the internal market?

Although the Draft AI Act does not refer to explainability in the context of transparency measures, it is implied in its Explanatory Memorandum. In particular, it notes that transparency should be established to ensure effective redress for those whose fundamental rights are infringed by an AI system, notwithstanding its compliance with other essential requirements.[70] To provide an effective remedy to those adversely affected by decisions based on the outputs of a high-risk AI system, users must be able to explain and justify to impacted individuals, in readily comprehensible terms, how and why the AI system reached its decision.[71] Notably, the amendments proposed by the European Parliament specify that transparency measures should enable users to comply with a new obligation to provide users adversely impacted by their systems, upon request, with a 'clear and meaningful explanation […] on the role to the AI system in the decision-making procedure, the main parameters of that decision taken and the related input data'.[72]

---

[69] Marco Tulio Ribiero, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?": Explaining the Predictions of Any Classifier' (Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016) <https://doi.org/10.1145/2939672.2939778> accessed 18 September 2023; Scott Lundberg and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions' (31st Conference on Neural Information Processing Systems, 2017) <https://doi.org/10.48550/arXiv.1705.07874> accessed 18 September 2023.

[70] Commission Proposal (n 9) s 3.5.

[71] Martin Ebers, 'Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s)' in Liane Colonna and Stanley Greenstein (eds), *Nordic Yearbook of Law and Informatics 2020-2021: Law in the Era of Artificial Intelligence* (The Swedish Law and Informatics Research Institute 2021).

[72] Amendments adopted by the European Parliament (n 9) Articles 13(1) and 68c(1).

Assuming that transparency requirements for high-risk AI systems will require that those systems be sufficiently transparent for the purposes of providing adversely impacted individuals with 'clear and meaningful' explanations, what is the technical threshold for such explanations? Different people require different explanations. This is based on several factors that are heavily context-dependent, such as one's socio-cultural background.[73] Although there are now several metrics used to measure the explainability of an AI system, this remains a nascent field where there is no reliable state-of-the-art.[74]

However, even if an AI system is sufficiently transparent for those overseeing the AI system to understand, explain, and justify its functionality, further value judgements will remain. How much human oversight is necessary? Is it required throughout the entire lifecycle of an AI system? How many humans are required for sufficient oversight? The Draft AI Act is silent on these points, other than in relation to high-risk AI systems used for biometric identification, which must be overseen by at least two people.[75] If empirical studies demonstrate that individuals are unable to provide reliable oversight of AI systems — for example, by unduly deferring to the AI system (known as 'automation bias') — should this be considered when determining what level of human oversight is sufficient?[76] Once again, the lofty normative aspirations of EU legislators may be thwarted when confronted with the nuances of reality.

---

[73] Hana Kopecka and Jose Such, 'Explainable AI for Cultural Minds' (Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction, 2020) <https://kclpure.kcl.ac.uk/portal/en/publications/explainable-ai-for-cultural-minds> accessed 18 September 2023.

[74] Mittelstadt (n 67).

[75] Commission Proposal (n 9) Article 14(5).

[76] Based on such empirical studies and the fact that human oversight is becoming a key mechanism for the governance of AI, some have argued that human oversight measures must be re-considered to ensure their effectiveness. For example, see: Ben Green, 'The flaws of policies requiring human oversight of government algorithms' (2022) 45 Computer Law and Security Review 105681

II. THE ROUTE TO STANDARDISING AI

The foregoing illustrates the normative challenges which the EU approach to AI standardisation poses to ESOs. In response, there is currently great scholarly interest in the possible approaches which ESOs will take to the development of AI harmonised standards under the Draft AI Act.[77] Although speculation is inevitable at this stage in the development process, there are at least two key features which are reasonably certain.

First, CEN and CENELEC will, where appropriate, use existing international standards as the basis for harmonised standards. Recital 3 of Standardisation Regulation and Recital 7 of the Draft Standardisation Request provide that ESOs should develop European standards in coordination with their international counterparts, including the International Organization for Standardisation (ISO) and the International Electrotechnical Commission (IEC). These inter-institutional relationships are governed by co-operation agreements, known as the Vienna Agreement (between CEN and ISO)[78] and the Frankfurt Agreement (between CENELEC and IEC).[79] These prohibit CEN and CENELEC from developing conflicting standards on the same subject matter

<https://doi.org/10.1016/j.clsr.2022.105681> accessed 18 September 2023; Johann Laux, 'Institutionalised Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act' (*Oxford Internet Institute*, August 2023) <http://dx.doi.org/10.2139/ssrn.4377481> accessed 18 September 2023.

[77] For example, see: Graeme Auld and others, 'Governing AI through ethical standards: learning from the experiences of other private governance initiatives' (2022) 29 Journal of European Public Policy 1822 <https://doi.org/10.1080/13501763.2022.2099449> accessed 18 September 2023; Laux, Wachter, and Mittelstadt (n 7); Pouget (n 57).

[78] Agreement on Technical Cooperation between ISO and CEN (the Vienna Agreement) (adopted 27 June 1991).

[79] Agreement on Technical Cooperation between CENELEC and IEC (the Frankfurt Agreement) (adopted 17 October 2006 June 1991).

unless they have 'particular needs', such as where international standardisation work does not exist or is deemed not to be sufficient or adequate for the fulfilment of the European Commission's standardisation request.[80] This means that, in practice, a significant proportion of harmonised standards originate as international standards developed by ISO and/or IEC, which are later adopted by ESOs, subject to any necessary modifications.[81] Given the maturity of the AI standardisation work carried out by ISO and IEC's joint technical committee, ISO/IEC JTC 1/SC42 (SC 42), relative to CEN and CENELEC's equivalent joint technical committee, CEN/CLC JTC 21 (JTC 21), the same is likely to be true in the context of harmonised AI standards.[82] Notably, the European standards mandated by the Draft Standardisation Request largely overlap with existing ISO/IEC standards or ones which are in advanced stages of development.[83]

---

[80] ISO and CEN, 'The Vienna Agreement — FAQs' (August 2016), 2 <https://boss.cen.eu/media/CEN/ref/va_faq.pdf> accessed 18 September 2023; CENELEC, 'CENELEC Guide 13 — Frequently Asked Questions on the Frankfurt Agreement' (edn 1, European Committee for Electrotechnical Standardization, June 2017), 4 <https://www.cencenelec.eu/media/Guides/CLC/13_cenelecguide13_faq.pdf> accessed 18 September 2023.

[81] According to CEN and CENELEC, out of 3,500 CEN and CENELEC harmonised standards cited in the Official Journal, 44% are based on international standards. See: CEN and CENELEC, 'CEN-CENELEC Response to the European Commission Standardization Strategy' (*CENELEC*, 8 August 2021), 7 <https://www.cencenelec.eu/news-and-events/news/2021/briefnews/2021-08-06-cen-clc_response_standardization_strategy_roadmap/> accessed 18 September 2023.

[82] SC 42 was established in 2017 and has since, as at 18 September 2023, published 20 standards with a further 31 under development, whereas JTC 21 has, as at 18 September 2023, published only two standards with 19 'projects' in progress. ISO, 'ISO/IEC JTC 1/SC 42 – Artificial Intelligence' <https://www.iso.org/committee/6794475.html> accessed 18 September 2023; CENELEC, 'CEN/CLC/JTC 21 — Artificial Intelligence' <https://standards.cencenelec.eu/dyn/www/f?p=305:22:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1E7E2C95DEE9A536E535BC6BAE2D4C821> accessed 18 September 2023.

[83] Micklitz (n 18) 23.

Second, those with relevant expertise in areas concerning fundamental rights will be involved, to some extent, in the standardisation process. This is required specifically in Annex II of the Draft Standardisation Request, and indirectly by Article 5(1) of the Standardisation Regulation which requires ESOs to 'encourage and facilitate the appropriate representation and effective participation of all relevant stakeholders', including 'social stakeholders'. Social stakeholders include civil society organisations focused on the fundamental rights implications of AI.[84] This obligation is strengthened in respect of civil society organisations that meet the criteria set out in Annex III of the Standardisation Regulation. These so-called 'Annex III organisations' can request funding from the EU to participate in the standardisation process and are able to participate in the standardisation process as observers.[85]

By further analysing these aspects of the European AI standardisation process, this Part suggests that ESOs are unlikely to provide answers to the hard normative questions outlined in Part II. The potential routes to standardising AI are likely to lead to a number of dead-ends for ESOs.

[84] For an example of such organisations, see the joint statement issued on 19 April 2023 by 151 civil society organisations calling on EU legislators to ensure that the AI Act puts fundamental rights first: Amnesty International, 'European Parliament: Make sure the AI Act protects people's rights!' (19 April 2023) <https://www.amnesty.eu/wp-content/uploads/2023/04/PDF-FINAL-Statement-European-Parliament-Make-sure-the-AI-act-protects-peoples-rights.pdf> accessed 18 September 2023.
[85] Standardisation Regulation, Article 16; CEN and CENELEC, 'Guide 25: The concept of cooperation with European Organizations and other stakeholders' (first published November 2021, updated on January 2023), para. 1.2.1 (Guide 25) <https://www.cencenelec.eu/media/Guides/CEN-CLC/cenclcguide25.pdf> accessed 18 September 2023.

A. International standards

During a webinar recently hosted by the AI Standards Hub, the Chair of JTC 21 suggested that a significant portion of harmonised standards would be based on existing ISO/IEC standards and specified certain existing standards which will be used.[86] Three of these specified ISO/IEC standards are analysed below.[87] The overall assessment is that, when it comes to hard normative questions, these standards are more descriptive than prescriptive. They thus do not provide the concreteness required for certification under the Draft AI Act.

(i) ISO/IEC TR 24027

The first standard considered here is a so-called 'technical report' on bias in AI systems and AI-aided decision-making.[88] According to ISO and IEC's internal regulations, technical reports are not permitted to include any expression which conveys: 'objectively verifiable criteria to be fulfilled and from which no deviation is permitted if conformance with the document is to be claimed' (defined as a 'requirement'); 'a suggested possible choice or course of action deemed to be particularly suitable without necessarily mentioning or excluding others' (defined as a 'recommendation'); or 'consent or liberty (or opportunity) to do something'

[86] AI Standards Hub, 'European AI standardisation in the context of the EU AI Act' (17 February 2023) <https://aistandardshub.org/events/european-ai-standardisation/> accessed 18 September 2023.
[87] This section does not aim to provide a comprehensive analysis of the existing ISO/IEC standards referred to in the webinar, but rather provides an analysis of certain standards for illustrative purposes. For analysis of other related ISO/IEC standards, see: Laux, Wachter, and Mittelstadt (n 7) 16-18.
[88] ISO and IEC, 'Information Technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision-making' (edn 1, *British Standards Institution*, November 2021) (Information Technology).

(referred to as a 'permission').[89] Given these formal constraints, the substance of technical reports is generally constrained to descriptive statements.

As the report begins, '[b]ias in AI systems is an active area of research' and, as such, limits its scope to describing 'current best practices to detect and treat bias in AI systems or in AI-aided decision making'.[90] While the report provides a definition of bias and its different forms, it refrains from defining what would constitute fair or unfair bias on the basis that fairness is a 'complex, highly contextual and sometimes contested' concept.[91] For example, the report notes that 'age-based profiling can be considered unacceptable in job application decisions.' At the same time, it suggests that 'age can play a critical role in the evaluation of medical procedures and treatment'.[92] Ultimately, the report simply provides an overview of potential sources of 'unwanted' bias in AI systems,[93] commonly used metrics for assessing bias and fairness,[94] and strategies which can be used to address bias throughout the lifecycle of an AI system.[95]

In this way, while the report describes the relevant methods for identifying and mitigating bias in AI systems, it prescribes neither when those methods should be used nor when action should be taken. Rather, it suggests that the relevant threshold should be derived from a combination of 'external requirements'. These can take the form of applicable legislation and official

[89] ISO and IEC, 'ISO/IEC Directives, Part 2: Principles and rules for the structure and drafting of ISO and IEC documents' (2021) (ISO/IEC Directives), para 3.1.8 <https://www.iso.org/sites/directives/current/part2/index.xhtml> accessed 18 September 2023.
[90] ISO and IEC, 'Information Technology' (n 88) vi.
[91] ISO and IEC, 'Information Technology' (n 88) sub-clause 5.3.
[92] ISO and IEC, 'Information Technology' (n 88) sub-clause 5.2.
[93] ISO and IEC, 'Information Technology' (n 88) clause 6.
[94] ISO and IEC, 'Information Technology' (n 88) clause 7.
[95] ISO and IEC, 'Information Technology' (n 88) clause 8.

guidance, as well as 'internal requirements', including an organisation's strategies and cultural values.[96] However, diverging interpretations of applicable legislation and different organisational cultures will not provide for consensus which can be readily standardised. This is particularly true in the case of AI, where there remains considerable normative and legal uncertainty, as illustrated in Part I. In the absence of stable regulatory rules or norms, the report risks leaving AI providers and users to determine normative thresholds which they find acceptable based on their commercial values and organisational interests.[97]

(ii) ISO/IEC TR 24029

The picture is much the same in respect of the second technical report analysed here, which focuses on assessing the robustness of neural networks.[98] Much like the first report, the substance of this second report is descriptive, owing to the novelty of the subject matter. In its introduction, the report notes that assessing the robustness of neural networks is 'an open area of research', and its scope is limited to a non-exhaustive list of existing methods of assessment used in industry, government, and academia.[99]

As a result, the report neither specifies when certain methods of assessment should be used nor what constitutes an acceptable level of robustness. It also acknowledges that there are limitations to testing and validating the robustness of neural networks given the 'inherent non-linearity of their behaviour'.[100] As discussed in Part I, it is very difficult, if not currently impossible, for AI developers

96 ISO and IEC, 'Information Technology' (n 88) sub-clauses 8.2.2 and 8.3.3.
97 Laux, Wachter, and Mittelstadt (n 7) 16.
98 ISO and IEC, 'Artificial intelligence (AI) – Assessment of the robustness of neural networks' (Edition 1, *British Standards Institution*, March 2021) (Artificial Intelligence).
99 ISO and IEC, 'Artificial Intelligence' (n 98) v and clause 1.
100 ISO and IEC, 'Artificial Intelligence' (n 98) sub-clause 6.1.

to reasonably foresee all plausible scenarios in which a system may not function as intended. While the report provides strategies and techniques to protect systems from adversarial attacks, it once again makes clear that these are not intended to be exhaustive and refers the reader to the 'literature' on this subject.[101] Like the first report, hard and normative questions are ultimately left unanswered.

(iii) ISO/IEC DIS 23894

The final standard analysed in this section is a guidance document on risk management systems for AI, which is currently a 'discussion draft'.[102] Unlike technical reports, guidance documents are permitted to contain recommendations, although they fail to provide requirements.[103] As such, the substance of the guide is limited to recommendations and descriptive statements on how organisations can or should address three key aspects of risk management: identification, evaluation, and mitigation.

In terms of identifying and evaluating risks, the guide provides a non-exhaustive list of stakeholders which 'should' be considered, including civil society organisations, individuals, and 'society'.[104] Other recommended considerations include whether an AI system can infringe human rights and the external expectations for that organisation's social responsibility.[105] To identify these stakeholders, rights infringements, and external expectations, the guide recommends the use of impact assessments.[106] While 'algorithmic impact

[101] ISO and IEC, 'Artificial Intelligence' (n 98) Annex A, A.1.
[102] ISO and IEC, 'Information Technology – Artificial Intelligence – Guidance on risk management' (*British Standards Institution*, January 2022) (Guidance on Risk Management).
[103] ISO and IEC, 'ISO/IEC Directives' (n 89) para 3.1.7.
[104] ISO and IEC, 'Guidance on Risk Management' (n 102) sub-clause 6.3.3.
[105] ISO and IEC, 'Guidance on Risk Management' (n 102) sub-clause 6.3.3.
[106] ISO and IEC, 'Guidance on Risk Management' (n 102) sub-clause 6.4.

assessments' may provide a promising approach to standardising the impact of AI systems on human rights, an important but ambiguous question for the assessor is determining which impacts are within scope.[107] On this point, the guide is less clear. It provides a non-exhaustive list of 'AI-related objectives' which risk management systems should facilitate (including accountability, fairness, robustness, transparency, and explainability). It then notes that 'the exact impacts will depend on the context in which the organization operates and the areas for which the AI system is developed and used'.[108]

In terms of mitigating identified risks, the guide recommends that measures are 'designed to reduce negative consequences of risks to an acceptable level, and to increase the likelihood that positive outcome can be achieved'.[109] Like the risk management requirements provided in Article 9 of the Draft AI Act, the guide does not elaborate on what constitutes an 'acceptable' level of residual risk. Once again, AI developers are left to make difficult value judgements which may, in certain contexts, have socially significant consequences.

---

[107] See, for example: Emmanuel Moss and others, 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest' (*Data & Society*, June 2021) <http://dx.doi.org/10.2139/ssrn.3877437> accessed 18 September 2023; Eve Gaumond and Catherine Régis, 'Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination' (*Lawfare*, 27 January 2023) <https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination#:~:text=Both%20HRIAs%20provide%20guidance%20to,unjustified%20infringements%20on%20human%20rights> accessed 18 September 2023.

[108] ISO and IEC, 'Guidance on Risk Management' (n 102) sub-clause 6.4.2.6 and Annex A.

[109] ISO and IEC, 'Guidance on Risk Management' (n 102) sub-clause 6.5.2.

B. Stakeholder expertise in fundamental rights

Some have argued that, in the context of AI standardisation specifically, broader and more effective participation from stakeholder organisations with fundamental rights experience can provide clarity where international standards are currently lacking.[110] The reason is that, because of a lack of stakeholder participation, international AI standards will not have the necessary expertise to provide adequate guidance for compliance with the Draft AI Act.[111] However, practice suggests something different: the ability of stakeholder organisations with fundamental rights expertise to participate effectively in the European standardisation process is not much better than at the international level.

First, although there is no equivalent to Annex III stakeholder participation in ISO/IEC, few organisations qualify as Annex III organisations in Europe. To qualify, a stakeholder organisation must be mandated by national organisations in at least-two thirds of EU Member States to represent stakeholder interests in the European standardisation process.[112] This condition is currently satisfied by only four organisations: the European Consumer Voice in Standardization (ANEC); the European Environmental Citizens' Organization in Standardization (ECOS); the European Trade Union Confederation (ETUC); and Small Business Standards (SBS).[113] While these organisations represent a broad range of interests, they still

[110] Galvagna (n 13) 24; Micklitz (n 18) 151.
[111] Galvagna (n 13) 23-24; Micklitz (n 18) 151.
[112] Standardisation Regulation, Annex III, sections 1(c), 2(c), 3(c) and 4(c).
[113] Kommission Arbeitsschutz und Normung, 'Annex III organizations: representatives of social stakeholders in European standardization activity' (*Kommission Arbeitsschutz und Normung*, April 2021) <https://www.kan.de/en/publications/kanbrief/4/21/annex-iii-organizations-representatives-of-social-stakeholders-in-european-standardization-activity> accessed 18 September 2023.

only represent a fraction of the fundamental rights and stakeholders implicated by the Draft AI Act.[114]

Second, alternative opportunities for other stakeholders to participate in the standardisation process are, in reality, very limited: stakeholders must either qualify as a 'liaison organisation' of the relevant ESO(s) or they must participate through their national standardisation organisation body.[115] CEN and CENELEC have published guidance on the requirements to qualify as a liaison organisation, which include an organisation having representatives in at least four CEN and/or CENELEC member countries. Those representatives must also be organisations rather than individuals.[116] In a recent survey of workshop participants whose organisations have relevant experience but are not involved in JTC 21, the Ada Lovelace Institute found that less than a quarter satisfied this requirement.[117] While stakeholders can participate through their national standardisation body, the fundamental purpose of providing Annex III organisations with financial and political support to participate in the standardisation process was because of the 'weak' representation of civil society in national standards bodies compared to big industry players.[118] Effective participation requires significant time and money, both of which civil society organisations generally have in short supply compared to industry players.

Third, stakeholder organisations which manage to participate in the standardisation process, including Annex III organisations, are ultimately limited in what they can do. For example, they have no voting rights and no right to

114 Galvagna (n 13) 24.
115 Galvagna (n 13) 26-29.
116 CEN and CENELEC, 'Guide 25' (n 85) para 2.3.
117 Galvagna (n 13) 27.
118 Parliament Resolution, 2010/2051(INI) of 21 October 2010 on the future of European Standardisation [2010] OJ C70 E/05, para 33.

require ESOs to either comply with their requests or to receive a written explanation why the ESOs will not comply.[119] As Micklitz observes, 'the only power [stakeholder organisations] have is knowledge and argument, but they have been given no tools to turn knowledge and argument into action in case they are outvoted or in case their comments are not respected'.[120] In other words, access to the standardisation process is not tantamount to effective participation in it.

But even if these obstacles did not exist or were less significant (in line with several recent proposals),[121] there are at least two key issues which mean that stakeholder participation in the standardisation process is unlikely to sufficiently address the normative challenges facing JTC-21. First, it is unclear how to guarantee that stakeholder organisations are representative of the public on whose behalf they claim to speak.[122] Second, perhaps the more fundamental issue is that ESOs would be on shaky legal ground if they were to make political decisions informed by stakeholder input. An example could be determining an appropriate metric for evaluating bias in AI systems and for what an acceptable level of bias would look like. In *Meroni v Highway Authority*,[123] the ECJ held that the delegation of discretionary power, as opposed to executive power, is not permitted because 'it replaces the choices of the delegator by the choices of the delegate, [and] brings about an actual transfer of responsibility'.[124] This case is now regarded as the

[119] CEN and CENELEC, 'Guide 25' (n 85) para 1.2.
[120] Micklitz (n 18) 177.
[121] For example, see: McFadden and others (n 18); Galvagna (n 13); Micklitz (n 18). See also: Annex III organisations, 'CEN and CENELEC's governance review in support of inclusiveness' (December 2022) <https://ecostandard.org/wp-content/uploads/2022/12/Annex-III-proposals-for-CEN-CENELEC-governance-review_Dec-2022.pdf> accessed 18 September 2023.
[122] For example, see: Justin Greenwood, 'Organized Civil society and Democratic Legitimacy in the European Union' (2007) 37 British Journal of Political Science 333; Micklitz (n 18) 102.
[123] Case C-9/56 *Meroni v Highway Authority* [1958].
[124] *Meroni* (n 123) 152.

benchmark for the lawfulness of private delegation of power by legal scholarship and the European Court of Justice (ECJ).[125] Moreover, according to *Meroni*, delegated power should be subject to the same conditions of control as those which it would have been if the delegator had exercised them.[126] The absence of this requirement — that is, the lack of judicial and administrative control — was a deciding factor in finding the delegation of power unlawful in *Meroni*. Whether this condition is now satisfied by the introduction of the Standardisation Regulation and the decision of the ECJ in *James Elliott Construction v Irish Asphalt*[127] is an ongoing area of debate in the literature.[128] While the former systematised the responsibility of the Commission to oversee the standardisation process, the latter established that harmonised European standards constituted 'EU law' and, therefore, could be subject to judicial review. Some, however, remain sceptical. For example, Veale and Zuiderveen Borgesius suggest that the value-laden nature of the Draft AI Act risks planting a 'constitutional bomb' under the NLF.[129]

[125] For example, see: van Gestel and Micklitz (n 28) 151; Case C-147/13 *Kingdom of Spain v Council of European Union* [2015], para 51.
[126] *Meroni* (n 123fr).
[127] Case C-612/14 *James Elliott Construction Limited v Irish Asphalt Limited* [2016] ECR 63.
[128] Megi Medzmariashvili, 'Delegation of Rulemaking Power to European Standards Organizations: Reconsidered' (2017) 44 Legal Issues of Economic Integration 353 <https://lup.lub.lu.se/record/76a75660-b022-4c63-b1f1-223f2aceaa84> accessed 18 September 2023.
[129] Veale and Zuiderveen Borgesius (n 39) 105.

Therefore, ESOs may be keen not to set it off by venturing into political decision-making.

III. THINKING OUTSIDE THE TECHNICAL STANDARDISATION BOX

As Part II suggests, answers to the hard normative questions which AI systems pose cannot come from within the European technical standardisation process. The implication is that, since ESOs do not have the necessary institutional authority, alternative actors or institutions which do have the requisite authority should be considered instead. Legislators and constitutional courts can be cited as clear examples. Following this line of thought, Laux and others propose a novel approach to European standardisation under the Draft AI Act, which they call 'ethical disclosure by default'.[130] According to this approach, ESOs would develop standards containing minimum disclosure requirements for AI providers. This, in turn, would enable 'local decision-makers with the legitimacy and knowledge' to determine ethical trade-offs and thresholds to access essential information.[131] While the aim of this approach is sound, this Part will demonstrate how the same is not true of its practical implications. This suggests that it may be necessary to think outside the technical standardisation box completely.

A. The faults of 'ethical disclosure by default'

The initial problem with the 'ethical disclosure by default' proposal is that it is unclear at what level minimum disclosure requirements should be pitched. Laux

[130] Laux, Wachter, and Mittelstadt (n 7).
[131] Laux, Wachter, and Mittelstadt (n 7) 22.

and others suggest that relevant tools for implementing these requirements would include: the application of bias and fairness metrics; transparency and explainability methods; model and data standardisation documentation; impact assessments; and '[a]ny other documentation describing ethical decisions made by providers or procedures used to make such decisions.'[132] Would harmonised standards, therefore, require the use of specific metrics, such as when testing AI systems for bias? If so, this would run into the same practical and legal obstacles already outlined in Part I and II.

Alternatively, minimum disclosure requirements could be more general. For example, standards could simply mandate bias testing but not specific metrics. In practice, however, this would allow AI providers to benefit from a presumption of conformity with essential requirements whilst simultaneously setting their own normative standards. To mitigate this, AI providers could be required to disclose information to relevant stakeholders upon request. This would include not only governmental organisations and judges but affected individuals and civil society organisations. The result would shift the setting of ethical decision-making away from AI providers. As such, there would be an extension of the disclosure requirements contained in the Draft AI Act. These are currently given to third-party auditors and public authorities charged with implementing and enforcing the Regulation. Laux and others endorse this approach but suggest that standards should set these additional disclosure requirements.[133] This is a curious suggestion since ESOs do not have the institutional competence to create new legal rights through standards. For this

[132] Laux, Wachter, and Mittelstadt (n 7) 23.
[133] Laux, Wachter, and Mittelstadt (n 7) 24.

proposal to succeed, it must be aimed at EU legislators who can amend legal requirements in the Draft AI Act, as others have done.[134]

Another issue with the 'ethical disclosure by default' approach is that, by focusing on shifting the setting of ethical decision-making away from ESOs and AI providers, it fails to acknowledge the importance of technical expertise in developing normative AI standards. Technical expertise is frequently required to explain the significance of how an AI system operates. Furthermore, ethical decisions can have technical implications which must be fully understood to ensure that they do not lead to further, unforeseen ethical problems. For example, limiting the data which AI developers can access to develop their systems can increase the risk of unreliable or biased outcomes. There may be privacy-preserving techniques which can mitigate this problem, as was discussed in Part I. However, again, awareness and understanding of those techniques rely on access to relevant technical expertise.

Finally, the 'ethical disclosure by default' approach does not provide a scalable solution for answering hard normative questions, or if it does, this is not elaborated on by those proposing it. Building on prior research,[135] Laux and others note that EU equality law and jurisprudence rarely offers '[s]pecific, measurable, and generalisable thresholds and trade-offs'. Therefore, equality requirements are given meaning on a case-by-case basis.[136] While this article has

---

134 Martin Ebers and others, 'The European Commission's Proposal for an Artificial Intelligence Act – A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 Multidisciplinary Digital Publishing Institute 589, 596 <https://doi.org/10.3390/j4040043> accessed 18 September 2023.

135 Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law and Security Review 105567 <https://doi.org/10.1016/j.clsr.2021.105567> accessed 18 September 2023.

136 Laux, Wachterm, and Mittelstadt (n 7) 23.

attempted to demonstrate how harmonised standards cannot provide generalisable metrics for determining whether the development or use of an AI system breaches equality law, there remains a need for compliance tools which improve the scalability and efficiency of reaching reasonable determinations. This is necessitated not only by the number of decisions which AI systems can make when they are in operation but also when the development of AI systems is fast-tracked in response to time-sensitive issues, such as the COVID-19 pandemic.

B. Standardisation inside the regulatory sandbox

In response to the deficiencies outlined above, this article proposes that greater attention should be paid to the role of another potential standardisation process: the development of regulatory conventions through regulatory sandboxing. In this context, a 'convention' can be understood as functionally equivalent to a rule-of-thumb; that is, a standard which is deemed to be reasonable but inconclusive. A notable example of this in a regulatory context is the 'four-fifths rule' which several US federal agencies, including the Equal Employment Opportunity Commission (EEOC), have used since 1978. This rule is relied on to help determine whether an employee selection procedure works to the disadvantage of members of a group with protected characteristics, such as race or sex (known as 'adverse impact').[137] More recently, the EEOC endorsed its application in the context of AI tools which employers are increasingly using to automate and inform selection processes.[138]

[137] EEOC, 'Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures' (2 March 1979) (Questions and Answers) <https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines> accessed 18 September 2023.
[138] EEOC, 'Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964' (18 May 2023) <https://www.eeoc.gov/select-issues-assessing-

According to the four-fifths rule, a selection procedure *may* have an adverse impact if the selection rate of individuals with protected characteristics is less than four-fifths (80%) the rate of the group with the highest rate for a selection procedure. Notably, violation of the 'rule' is not conclusive evidence of unfairness but rather potential evidence of it, warranting further enquiry. As the EEOC clarifies, failure to satisfy the four-fifths rule 'merely establishes a numerical basis for drawing an initial inference and therefore requiring additional information'.[139] There are several other factors which must be considered to determine whether a selection procedure produces an adverse impact, such as the size of the applicant pool and whether the employer discourages certain applicants from applying in the first place. In this way, the four-fifths rule respects the need for subjective judgement when determining ethical thresholds and trade-offs, such as in the context of equality law, while simultaneously providing a scalable tool which can streamline those assessments.

However, scalability is only half the problem. Regulatory conventions must also be reliable indicators of risk. As illustrated in Part I, a significant challenge to developing AI standards which can reliably identify, assess, and mitigate risks to fundamental rights is defining the intended use(s) of complex, general-purpose AI systems. This is because they often respond to new data in unexpected and unpredictable ways. While the EU regulatory approach has so far failed to engage meaningfully with how these challenges should be met, it nonetheless provides a promising means of doing so: regulatory sandboxing.

---

adverse-impact-software-algorithms-and-artificial-intelligence-used?utm_content=&utm_medium=email&utm_name=&utm_source=govdelivery&utm_term=> accessed 18 September 2023.

[139] EEOC, 'Questions and Answers' (n 137).

Article 53(1) of the Draft AI Act suggests that AI regulatory sandboxes will be established in Member States and/or at the EU level in order to provide a controlled environment in which AI systems can be developed and tested under the supervision of a competent authority. This is done with a view to ensuring that they comply with applicable legal requirements before they are placed on the EU market. This would provide an opportunity for AI providers, regulators, and other stakeholders to see how AI systems are applied in, and respond to, different contexts without testing them on the society at large. In doing so, relevant stakeholders could test and develop tools to identify, assess, and mitigate risks which materialise. Through extensive, controlled experimentation, lessons could be learnt and best practices could emerge. These could, in time, translate into regulatory conventions that reflect regulators' reasonable expectations of how AI systems will comply with applicable legal requirements. This would be an iterative process, whereby regulatory guidance and conventions informed by regulatory sandboxing would be applied in future regulatory sandboxes. Depending on the results, they could be further calibrated to ensure they remain reliable indicators of identified risks. This process would be heavily collaborative and iterative, drawing on the different competences and perspectives of the AI industry, regulators, and other stakeholders.[140]

At this stage, it is worth clarifying several points in anticipation of potential criticisms of the above proposal. First, although regulatory conventions alone would not provide the same level of legal certainty as harmonised standards - because they function as indicators of potential, rather than conclusive, evidence

[140] For a recent report on the use of regulatory sandboxes as a means of regulatory experimentation in the context of AI, see: OECD, 'Regulatory Sandboxes in Artificial Intelligence', (OECD Digital Economy Papers No 356, July 2023) <https://doi.org/10.1787/8f80a0e6-en> accessed 18 September 2023.

of unacceptable risks - they could achieve the same level if it is possible for AI providers to test their results with regulators, including through regulatory sandboxing. Such an outcome would be consistent with the stated aim of regulatory sandboxing under the Draft AI Act, which is to ensure that AI systems are compliant with applicable legal requirements before being placed on the market. Second, regulatory sandboxing inevitably focuses on particular use cases in specific industries, which contrasts with the horizontal approach to regulating AI systems under the Draft AI Act. As already suggested, this should be seen as a positive feature of regulatory sandboxing in this context, where the EU regulatory approach has so far failed to link the normative demands of its legal requirements with their practical application through use cases. However, it would be a mistake to assume that lessons in one context cannot inform lessons in another. AI raises many of the same questions across different industries. Although the answers in each context may be different, regulatory conventions which emerge in one environment may inform conventions in another by identifying relevant similarities and distinctions.

CONCLUSION

This article has attempted to widen the debate on the role of standards under the Draft AI Act specifically, with an eye to future equivalents. The development and use of AI systems raises numerous difficult normative questions. It is important to think carefully about how these questions should be answered in practice and by whom.

EU legislators have so far avoided responsibility for answering these questions by failing to translate the normative aspirations of EU AI policy into

clearly defined legal requirements in the Draft AI Act. ESOs who are, in turn, relied on to operationalise these requirements, are left to make sense of two conflicting demands from their political superiors. On the one hand, ESOs are not authorised to make political choices when developing harmonised standards to operationalise legal requirements. On the other hand, to operationalise essential requirements under the AI Act, as mandated, political choices must be made.

Rather than trying in vain to square these conflicting demands within the confines of technical standardisation, it is time to think outside the box and consider whether there are more viable means of navigating the hard normative questions which the development and design of AI systems inevitably raise. While reducing contested concepts and ethical trade-offs to objectively verifiable criteria and precise technical solutions may be wishful thinking, the reality is that those developing and using AI systems need scalable tools to help ensure that those systems align with societal and regulatory expectations. As a form of pragmatic compromise, this article proposes the development of regulatory conventions. These are meant to represent regulators' reasonable expectations of which metrics and what thresholds should be applied in areas of regulatory ambiguity, such as non-discrimination. Critically, these conventions should be developed through regulatory sandboxing, where expertise of the AI industry, regulators, and other relevant stakeholders can be combined in a complementary manner. By following this approach, the hard normative questions facing the AI industry and its regulators could become a little easier to answer.