Can we standardise the frontier of AI?

Huw Roberts,[1*] Marta Ziosi[2]
[1] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK
[2] Oxford Martin AI Governance Initiative, 34 Broad St, Oxford OX1 3BD, UK

*Email for correspondence: huw.roberts@oii.ox.ac.uk

Abstract

International standards have been promoted as a mechanism that can support the governance of advanced AI through explicating national-level regulations, supporting interoperability between different jurisdictions, and guiding best practices. However, there is significant ambiguity about what should be standardised for advanced AI systems, when, and by whom. In this paper, we explore the possibilities and limitations of international standards as a governance tool for advanced AI. Given the breadth of issues standardising advanced AI covers, we focus on the case study of standards designed to address fairness and bias in large language models (LLMs) and use it to draw broader theoretical insights. To explore this case, we conducted 50 interviews with AI standards experts to understand how the institutional competencies of standards development organisations (SDOs) and the characteristics of advanced AI influence the efficacy of international standards as a governance tool. Interviewees highlighted that international SDOs have a strong reputation, robust processes, broad international representation, and track record of producing impactful standards. However, they face challenges regarding adoption, participation, governing value-laden issues, and the speed and complexity of technological development. Based on this, we argue that traditional SDOs should not "standardise the frontier" *per se* and should instead focus on governing well-established "trailing-edge" issues, while developing regulatory intermediary partners to address "leading-edge" technical issues.

Keywords: artificial intelligence, technical standards, fairness, governance, institutions

1. Introduction

Late 2022 and early 2023 saw the mass commercialisation of advanced artificial intelligence (AI) systems,[1] like OpenAI's ChatGPT. Attempts to govern these technologies are already underway, with governments publishing AI regulations (Roberts et al., 2023) and industry producing Frontier AI Safety Frameworks (FMF, 2024). International standards can play an important role in the governance ecosystem by explicating regulations, supporting international interoperability, and providing companies with best practices (Cihon, 2019; Kerwer, 2005). However, there remains significant ambiguity surrounding which aspects of advanced AI should be standardised, when, and by whom (Schuett, 2024). This creates difficulties for realising the benefits standards could offer.

This paper examines the role international standards can play in governing advanced AI, based on the institutional competencies of standards development organisations (SDOs) and the characteristics of advanced AI (Koremenos et al., 2001). Given the breadth of issues that standardisation for advanced AI covers, we focus specifically on the case study of standards designed to address harmful biases in large language models (LLMs).[2] While LLMs are just one type of advanced AI, and harmful bias a single governance issue, utilising this case study illustrates many of the issues involved with standardising advanced AI. The case study approach is not designed to be representative or generalisable across all areas, but highlights several technical, ethical, and institutional considerations associated with standardising advanced AI.

To address this topic, we conducted 50 semi-structured interviews with AI standards experts. From these interviews, we found that international SDOs, particularly the "Big Three",[3] have strong reputations, robust processes, broad international representation, and track record of producing impactful standards, suggesting they can play some role in governing advanced AI. However, they face challenges regarding adoption, participation, governing value-laden issues, and the speed and complexity of technological development.

From assessing this case study, we draw several insights about the role of international standards in governing advanced AI. First, we highlight how, unlike many products or mature technologies, technical standards for advanced AI should not be seen as the lowest level of the governance stack, with technical specifications also needed to explicate these documents. Second, we argue that the institutional competencies of SDOs mean they are not the best institutions to be developing granular technical specifications. Instead, standards bodies are suited to addressing higher-level "trailing-edge" governance problems that are more well-established. Thirdly, we argue that SDOs should develop "regulatory intermediary" (Abbott et al., 2017) partnerships with newer and more agile institutions that are developing "leading-edge" technical specifications. This will support a clearer institutional division of labour, strengthen harmonisation, and lessen opportunities for "forum shopping" by states and private actors (Randall Henning & Pratt, 2023).

[1] Advanced AI systems are machine learning-based technologies, such as Large Language Models and multimodal AI systems, capable of performing complex tasks like language understanding, content generation, and decision-making at scale.

[2] LLMs are a type of advanced AI system with an extensive number of parameters, designed for natural language processing tasks, like text generation and summarisation.

[3] The "Big Three" – ISO, IEC, and ITU – are the world's principal standards-making bodies that are formally recognised by the WTO.

The remainder of this article is structured as follows. Section 2 introduces international standards, SDOs, and their work on advanced AI. Section 3 provides an overview of the case study of fairness in LLMs. Section 4 outlines the methodological approach taken in this paper. Section 5 presents the findings from our interviews. Section 6 considers the implications of the findings.

2. Technical standards for advanced AI

There is no agreement regarding how "international standard" should be defined. According to two of the most prominent SDOs – the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC) – a standard is,

> "A document, established by *consensus* and approved by a *recognized body*, that provides, for common and repeated use, rules, guidelines, or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context" (ISO/IEC Guide 2:2004).

Understood in this narrow sense, international standards are technical documents developed by international SDOs, like ISO, IEC, and the International Telecommunication Union (ITU), that have formal procedures for reaching consensus and whose work is widely recognised by national governments and international organisations.

Some scholars take a more encompassing view of the term. For example, Büthe and Mattli, (2011, p. 17) define international standards as "rules established by expert bodies prescribing de jure or de facto the quality or performance of a given practice, procedure, or product." This definition includes initiatives developed by informal institutions, like business consortia and open source communities, that have an international impact. However, such a broad definition is unhelpful for this paper because it encompasses a huge diversity of governance documents, prohibiting a targeting analysis. As such, international standard is understood in the paper in line with the definition provided by ISO/IEC.

2.1. What standards do

Technical standards are generally developed to deal with four categories of governance problems: technological interconnectivity, transactional interconnectivity, physical externalities, and policy externalities (Abbott & Snidal, 2001). Technological interconnectivity standards facilitate technical interoperability and without these standards, technologies would not function together. For example, trying to connect pieces of railway track with different size gauges would result in a route that is unusable for trains. To date, there has been comparatively little demand for technological interconnectivity standards for advanced AI systems as, aside from open source models, relatively little technical interoperability is needed between different companies (i.e., AI is not like a rail gauge). Nonetheless, some standards have been developed which facilitate model, data, and cloud interoperability, like ISO/IEC 19941:2017, or to facilitate the integration of AI into networks, like ITU-T Y.3176. The second type of interconnectivity standards are transactional interconnectivity standards. They focus on facilitating transactions and interoperability between organisations, such as through standardised contracts and procedures. These standards are not a technological necessity but simplify bargaining. The few cases of transactional interconnectivity standards for AI can also be found in the open-source community, notably through responsible AI licences which are used for specifying the terms under which models and datasets are shared (McDuff et al., 2024).

The other categories of problems standards attempt to address are physical and policy externalities. Physical externalities, like pollution, occur when an actor's behaviour directly affects another. The goal of physical externality standards is to optimise the level of externality-generating activity, such as by introducing safeguards. Policy externalities occur when laws or policies in one jurisdiction affect actors in another (Abbott & Snidal, 2001), as is this case with the so-called "Brussels Effect" from many EU regulations which leads companies to comply with EU rules even when they are operating outside the bloc (Bradford, 2020). These standards mitigate policy externalities by facilitating regulatory interoperability and promoting uniform compliance.

Significant effort has been made to develop standards focused on mitigating physical externalities (i.e., making AI safer and more ethical) and policy externalities (i.e., focused on strengthening interoperability between different jurisdictions). Indeed, many international standards attempt to do both through developing or explicating safety or ethics governance frameworks that can be commonly used across jurisdictions. Standards focused directly on mitigating physical externalities include safety standards relating to information security (ISO/IEC 27001:2013) and functional safety of AI systems (ISO/IEC TR 5469:2024).

2.2. How standards are used and by whom

Interconnectivity standards are developed by industry actors to support interoperability and require little incentivisation. This is because companies often recognise benefits in aligning their products and services, which facilitates broader market access and reduces compatibility issues (David & Steinmueller, 1994). In contrast, international standards addressing physical externalities typically need to be incentivised by regulatory, reputational, and commercial mechanisms.

Governments may directly reference standards in legislation, effectively transforming voluntary standards into legal obligations (Coglianese, 2023). This practice is widespread and allows regulators to leverage the technical expertise of SDOs without having to recreate complex specifications themselves, in effect using them as "regulatory intermediaries" (Abbott et al., 2017). In the context of AI, technical standards are being utilised to support compliance and enforcement, including in the EU, where standards developed by the European standards organisations CEN and CENELEC will create a presumption of conformity with the EU AI Act (Cantero Gamito & Marsden, 2024). Many of these standards derive from international standards developed by ISO/IEC (Soler-Garrido et al., 2024).

Certification schemes serve as another important driver of standards adoption. A company may be motivated to achieve certain standards (e.g., ethics or safety standards) if they can use certification to demonstrate their achievements to customers or other relevant actors (Christmann & Taylor, 2006). Several sectors are subject to a "New Walmart Effect", whereby downstream "lead firms" who have a significant market share demand certification and through this, change the behaviours of upstream suppliers (Vandenbergh, 2007). However, because of the general-purpose characteristics of AI and in turn, the use of models across sectors, the power of downstream lead firms is more diffuse, posing potential challenges to voluntary certification schemes.

International trade dynamics also play a role in standards adoption as under the WTO's Technical Barriers to Trade (TBT) Agreement, countries are encouraged to base their technical regulations on international standards where they exist (WTO, 1995). However, Article 2.4 of the TBT provides broad exceptions where such standards would be ineffective or inappropriate for achieving legitimate policy objectives, which leads to standards divergence in practice (Charnovitz, 2005).

3. Case study: LLM fairness

Given the diversity of technologies, policy issues, and sectors standardisation for advanced AI covers, it would be futile to try and assess the complete landscape. Instead, we focus here specifically on the case study of fairness standards for LLMs and use it to draw broader theoretical insights. We selected the case of fairness because it is a complex, yet relatively mature policy area in the field of AI governance, which has been complicated by the latest generation of AI technologies (Mökander et al., 2023). We chose to focus on LLMs because they are currently the most prevalent type of advanced AI.

3.1. Bias in LLMs

Harmful bias in algorithmic systems is nothing new (Friedman & Nissenbaum, 1996), with numerous examples of task-specific AI systems designed to perform specific functions like ranking job applicants (Dastin, 2018) or diagnosing illnesses (Obermeyer et al., 2019) demonstrating biases. These biases stem from design and deployment decisions, including data selection, model calibration and optimisation, flawed human oversight, and deployment in inappropriate contexts (Danks & London, 2017).

LLMs are trained on general data, used downstream across multiple contexts, and undergo frequent updates, which exacerbates problems of AI bias and complicates effectively identifying and mitigating these biases (Bommasani et al., 2022). LLMs can produce "intrinsic biases" because datasets used to train the models are often scraped from the internet and reflect historical and societal inequities, prejudices, and stereotypes (Ethayarajh et al., 2019). The broader data used for training LLMs means they are more exposed to societal biases than task-specific AI, which are only trained on task-relevant data. They also produce "extrinsic biases" based on the interactions between the user and the system, such as through prompting (Goldfarb-Tarrant et al., 2021).

LLMs are foundation models, meaning any intrinsic biases will be repeated downstream in systems that build on the model, leading these biases to cause harm on a far larger scale (Mökander et al., 2023). The impact of bias is not limited to a single use case (e.g., a single company's recruitment algorithm), but affects all individual and commercial users of an LLM. LLMs are also increasingly multimodal, creating a potential for bringing together biases coming from text (e.g. biased word embeddings) as well as image data (e.g. stereotypical representations).

These harmful biases can lead to representational harms arising from the perpetuation of harmful attitudes towards a social group (Gallegos et al., 2023), including their misrepresentation, stereotyping, or disparate system performance. For example, disability representations encoded in language models often inadvertently perpetuate undesirable social biases (Hutchinson et al., 2020). Biases can also lead to allocative harms, which arise when resources are unjustly allocated or re-allocated, due to a skew in model output. This includes lost opportunities or direct and indirect discrimination (Gallegos et al., 2023). For example, an LLM being used to filter resumes for software engineers may prefer applicants with stereotypically male names because they have historically been hired more frequently (Dastin, 2018). How biases play out across languages varies. Moreover, some models extend representations and structures that are particular to a – typically majority – language to other languages, often with unintended consequences.

3.2. Making LLMs "fairer"

To address these harmful biases, LLMs must produce outputs that are "fair", but what constitutes "fair" is highly contentious.[4] Broadly speaking, fairness can be understood as the just treatment of individuals or groups, stemming from either the process through which a decision is made or the distribution of a good against an expected baseline (*Bias Review*, 2020). However, the contested and contextual nature of what constitutes a fair procedure or outcome means there is not a universal formula that can be applied to make LLMs fair (Binns, 2018). Discrimination law provides requirements against which to check an AI system's performance. Yet, laws on discrimination vary widely across jurisdictions. Additionally, they might only cover a narrow understanding of fairness, leaving out aspects of bias which are not strictly illegal but could still be harmful to minority populations (e.g. harmful stereotyping). Fairness interventions in LLMs may also require trade-offs with other goals, such as accuracy or efficiency, which could impact an AI system's performance (Whittlestone et al., 2019).

Against this backdrop, international SDOs have published or proposed technical standards that can directly or indirectly support "fairer" LLMs. This work has predominantly been undertaken by ISO and IEC through their joint technical committee and the IEEE. The ITU's AI standards have focused more on the integration of AI in networks, with fairness only mentioned in passing in sector-specific standards, such as standards on AI for Health (ITU, 2022).

Existing standards which can support fairness in LLMs generally fall into three categories. First, there are foundational standards, like ISO/IEC 22989:2022, which defines key concepts related to fairness (e.g., section 5.15.9. "AI bias and fairness") and is being updated to include terms related to advanced AI. These standards provide a common conceptual basis for governance. Second are standards focused specifically on evaluating or mitigating fairness and bias. IEEE 7003-2024 provides processes and methodologies to help users address bias, while IEEE P3419 will establish criteria for evaluating LLMs. ISO/IEC TS 12791:2024 discusses sources of bias in an AI system, statistical metrics to assess bias, and methods for treating unwanted bias. However, this standard does not specifically focus on bias in LLMs, raising questions over its applicability. Finally, there are broader governance standards, like ISO/IEC 23894:2023, which outlines guidance on risk management for AI and includes provisions related to harmful biases. Such standards are important to ensure that biases are not solely considered as a priority in technical terms, but also as an important part of organisational

[4] We focus here on fairness relating to use of these systems rather than broader structural questions, like unequal access to technologies.

processes. This outline is designed to be indicative of the work taking place in SDOs, not an exhaustive list.

4. Methodology

To understand the strengths and limitations of international standards for governing advanced AI, we conducted semi-structured interviews with AI standards experts. A purposive, maximum variation sampling approach was used to capture the diverse perspectives of those developing standards in traditional SDOs, competing institutions like business consortia, and standards adopters in government and the private sector (Palinkas et al., 2015). We focused on capturing a diverse set of opinions across the standards-making and -taking landscape to identify patterns that cut across cases. This was necessary as while standards-makers possess a deep knowledge about institutional processes, they also have vested interests in their institution developing standards. As one interviewee put it, "to some standards-makers, everything looks like a nail that needs to be hammered by the ISO" (Interviewee 25). This highlights the need for the perspective of standards-takers, as well as those developing competing initiatives outside of traditional SDOs.

Because we utilised a maximum variation sampling approach, we interviewed a larger than average sample of 50 experts (Dworkin, 2012). This is because reaching practical saturation (i.e., when further interviews produce little additional information for substantively answering the research question) required a larger number of participants than in a more targeted sample (Guest et al., 2006). For the questions, we asked interviewees about the current AI standards landscape, the applicability of these standards to advanced AI, and the appropriateness of SDOs for filling gaps compared to other types of institution.

Conducting interviews – rather than just undertaking a theoretical analysis of institutional competencies and technological characteristics – is necessary because of the opaqueness of many aspects of standards-making and -taking. For example, there is very little publicly available data on the standards-making process, including for many institutions, which organisations or individuals proposed a standard and took part in its drafting. Similarly, there are no public datasets on who is adopting which standards. It was thus necessary to speak with insiders to understand how SDOs, markets, and regulators are responding to the opportunities and challenges posed by advanced AI.

5. Can standards make LLMs fairer?

There was general agreement among our interviewees that international standards could play some role in making LLMs fairer. Interviewees stressed that the established technocratic procedures of the SDOs active in standards-making for AI provide them with a generally strong reputation among governments and industry. However, four key challenges were identified, which we outline below.

5.1. Adoption incentive problem

Interviewees highlighted that some international standards, like the AI management standard ISO/IEC 42001, are beginning to be utilised (Anthropic, 2025; Duffer et al., 2024; Infosys, 2024). However, this is the exception rather than the rule, with interviewees stressing that incentive problems led to low adoption rates of international AI standards.

One reason provided for the lack of take-up is because there is often overlap in the content of international standards produced by SDOs and standards-like documents produced by businesses consortia and government institutions. In recent years, industry consortia such as the Frontier Model Forum (FMF), engineering consortia such as MLCommons, and government entities like the AI Safety or Security Institutes have been producing guidance (Buhl et al., 2025), benchmarks (MLCommons, n.d.; UK AISI, 2025), and technical reports (FMF, 2025) that often function similarly to formal technical standards in practice. These bodies tend to be more agile than traditional SDOs and develop their documents in closer collaboration with advanced AI companies.

The fragmentation of governance across multiple institutions can increase flexibility and enhance overall expertise. However, it can also lead to inconsistency of rules and obligations, create opportunities for forum shopping, and ambiguity over which rules one should adopt (Randall Henning & Pratt, 2023). While some of these documents are complementary, participants stressed a general sense of being overwhelmed by the number of options.

Interviewees considered two pathways that could support greater adoption. First, mandating the use of standards through regulation. Steps are being taken in some jurisdictions to achieve this. Notably, in the EU, the European standards organisations CEN and CENELEC have been delegated responsibility for explicating key details of the legally binding AI Act. These standards are not legally mandatory, but they can be used to demonstrate conformity, indicating that they will receive significant take-up. However, while it has been confirmed that some ISO/IEC international standards will be directly adopted by CEN-CENELEC, in other areas, international standards were deemed inadequate for explicating the AI Act (Soler-Garrido et al., 2024). This suggests that some degree of divergence will emerge between voluntary international standards and *de facto* mandatory European standards. In particular, there is significant ambiguity as to whether a "Brussels Effect" will emerge, whereby companies use EU AI standards in third countries (Bradford, 2020).

Second, interviewees argued certification schemes could encourage adoption, but stressed several issues were currently present. Participants noted the presence of multiple uncertainties when it came to certification due to LLMs constantly changing, meaning certification might lose its value shortly after it is issued, and that organisational certification may be more desirable. They also noted that contractual obligations could incentivise standards uptake, with upstream providers potentially facing demands from downstream clients. However, as discussed above, the concentrated market dynamics (Korinek & Vipra, 2025) and the general-purpose nature of advanced AI systems, means "lead firms" who can pressure voluntary certification are more diffuse (Bartley, 2022).

5.2. Standardising value-laden issues

Standards bodies are technocratic institutions, with disagreements in the standards-making process needing to be laid out on technical grounds. For many types of standards, this is appropriate and helps avoid an overly politicised process; however, given the sociotechnical and value-laden nature of fairness, many participants were sceptical about whether standards bodies are equipped to deal with this type of issue. One participant stressed the lack of a tradition of dealing with value-laden topics and issues in SDOs. They stated,

> *"Standards come from a space that was originally all about purely technical things that you can say. Here is the measurement. Here is the metric. This is what the device needs to look like. This is how we measure whether it is safe or not, et cetera. Those kinds of things. But if you are talking about fairness, it becomes much more of a question of well under which circumstances? The social aspect is not something that the standardisation ecosystem has traditionally been engaged with."* (Interviewee 12)

The participant explained how the focus of such processes tends to be on deciding about what constitutes a "proper procedure" rather than of resolving substantial ethical matters. Others were more optimistic about the role SDOs could play, stressing that they should not be prescribing specific ethical positions, but could offer technical guidance at a higher level of abstraction. For example, one international standard – ISO/IEC TR 24027:2021 – conceptualises bias as a special case of an accuracy problem and does not deal directly with the concept of "fairness". Accordingly, questions remain as to whether a standard is the right format for presenting complex moral discussion.

Other participants questioned whether a governance issue as contextual as fairness could be standardised at an international level because of differing value systems and equalities laws. This is particularly salient because of the process of SDOs where consensus must be reached among diverse stakeholders with potentially conflicting ethical frameworks (Lewis et al., 2020). One participant noted that a standard following the US's four-fifths rule for fairness[5] would be illegal to deploy in the UK because of the country's equalities laws. These complexities are inherent challenges in governing advanced AI systems, like LLMs, because of the autonomy they add and the value-laden nature of their outputs.

5.3. Participation questions

There was some positivity in respect to both the geographical and sectoral diversity of participants involved in AI standards-making. For example, ISO/IEC JTC 1/ SC42 has a broad national participation, with 45 participating member countries and 25 observing members from around the world. One interviewee stated that AI was the most diverse standards-making process that they had been involved in. Nonetheless, three groups were highlighted as being underrepresented in the standards-making process due to the financial and time costs of meaningfully engaging in standards-making: civil society organisations, small- and medium-sized enterprises, and Global South stakeholders. It was stressed that the complexity of procedures means that those who are not heavily involved and knowledgeable could be easily sidelined in discussions. One participant estimated that it would take over 18 months for someone to become comfortable being involved in the standards-

[5] The four-fifths rule states that if a protected group's selection rate is less than 80% of the most favored group's rate, it may indicate potential discrimination.

making process. Interviewees highlighted that Global North members came with far larger delegations and were more active in discussions. In contrast, Global South members often abstained from voting on issues because of their lower levels of participation. Without meaningful participation of these three stakeholder groups, it is impossible to develop "fair" and representative standards that reflected the interests of citizens globally.

In contrast, multiple interviewees highlighted that certain established Big Tech companies were wielding a significant influence in the standards-making process. As one interviewee put it,

> "*The strongest asymmetry is between those spending 80 hours a week in standardisation, who are getting paid a massive amount by Big Tech to support this, and experts who are representing civil society and struggle to put five hours a week in and attend meetings.*" (Interviewee 29)

Some interviewees pointed to the ability of these companies to leverage their transnational operations (ten Oever & Milan, 2022) through seeding experts into multiple national standards committees. This allows these companies to influence the positions of national committees and in turn, international standards, because of the one country, one vote system in ISO/IEC.

However, it was also emphasised that this active engagement did not extend to all Big Tech companies. Several advanced AI companies, like OpenAI, were said to be largely absent from the standards-making process because they have no tradition of being involved. It was suggested that they have little time and incentive to upskill in SDO standards-making processes because unlike other technologies, such as the Internet, which require technological interconnectivity standards for their functioning (Yates & Murphy, 2019), international AI standards primarily designed to address externalities are not technologically necessary. As such, these companies are generally less invested in SDOs and are instead partaking in governance work in more agile institutions which are creating standards-like documents.

Efforts are being made to address this participation problem. For instance, both NIST and the AI Standards Hub have established pilot initiatives to invite a broader range of stakeholders to comment on drafts of standards, with this then being fed back into the standards committees (AI Standards Hub, 2023; NIST, 2024). Nonetheless, some civil society actors voiced their scepticism about such initiatives, stating anything short of full participation would not ensure meaningful influence.

5.4. Speed and complexity of technological development

The final theme of our interviews relates to the speed and complexity of technological development taking place and the difficulties SDOs' institutional procedures face in adapting to this. Mitigating harmful biases in "task-specific" AI systems (i.e., systems designed to address a specific problem) that deal with "classification" tasks (i.e., systems which categorise data into groups) is well-researched, with several techniques developed, such as data preprocessing to remove or balance biased data, algorithmic adjustments to reduce discrimination during model training, and post-processing methods to correct outputs. The specific nature of the data, task, and context allows for techniques to be developed or tailored to particular bias problems. As discussed in Section 3.1, the size of LLMs, the frequency of model updates, and the complex relationship between the LLM and the downstream task-specific applications that build upon it also make identifying and mitigating biases extremely challenging (Mökander et al., 2023).

Alongside this, LLMs output text rather than classifications, making the idea of "fair" outcomes more complex. The problem is not just determining whether the number of a specified good (e.g., job interviews) is distributed according to an expected baseline. Instead, it requires that various individuals or groups are not unfairly disadvantaged across various distributions within the worldview of the LLM. This problematises the applicability of many existing standards that were largely developed with task-specific, classification systems in mind. It also means that there is less well-established consensus on how fairness for LLMs can be standardised, which is problematic for SDOs where consensus is necessary for standards to pass.

Interviewees also highlighted that SDOs' processes are ill-equipped to address cutting-edge AI issues. ISO standards take 18, 24, or 36 months to develop, which indicates that traditional SDOs are not sufficiently agile for addressing many governance problems related to LLMs and fairness. Some SDOs have special procedures to speed up standards-making, like ISO/IEC's Publicly Available Specifications (PAS) which allows external organisations to develop their own technical specifications and present them for approval under ISO's framework, and fast track procedures which expedite processes. Yet, these procedures are still slow by AI development timelines and they create risks of corporate capture through less rigorous scrutiny. As AI continues to advance rapidly, with new emergent capabilities (Berti et al., 2025) and the development of agentic AI (Schneider, 2025), SDOs' processes to reach consensus will likely increasingly be put under strain.

Problems with the slowness of institutional procedures are compounded by the general-purpose and sociotechnical nature of LLMs, which means industry stakeholders from a variety of sectors, as well as civil society actors, have an interest in how these technologies are governed. Many of these stakeholders do not have an established tradition of being involved in standards-making. Efforts to ensure broad participation are desirable; however, some participants stressed that efforts to make the standards-making process more inclusive has had an unintended consequence of further slowing down the standards-making process. This is because many participants were unable to consistently take part in the process, meaning new members joined discussions and reopened topics where consensus had previously been nearly reached. It is also due to new members not fully understanding aspects of SDO procedure, leading them to engage improperly.

6. Can we standardise the frontier of AI?

The case study of LLM fairness is illustrative of the wider opportunities and challenges SDOs face in standardising the frontiers of AI. The speed of AI development, combined with the time needed to reach formal consensus and the value-laden nature of standards suggests traditional SDOs will be ill-equipped to develop "leading-edge" AI standards. Nonetheless, international standards can still play a role in establishing higher-level governance rules for more established issues related to advanced AI.

6.1. "Standards" vs. "specifications"

Governance functions at different layers. At a high level, institutions produce soft law mechanisms, like high-level governance principles, or hard law mechanisms, like regulation. However, these documents are typically not specified at a sufficiently granular level to be useful for organisations trying to demonstrate compliance with law or follow best practice. Consequently, "regulatory intermediaries" are often turned to who can provide further guidance for facilitating implementation or monitoring compliance (Abbott et al., 2017). SDOs are a typical example of a regulatory intermediary which produce lower-level governance documents that can be used be the target of regulation or best practice.

For traditional products, further layers of governance are not necessarily needed below standards. For example, a standard outlining the dimensions and performance requirements for a shipping container is sufficient for development and testing. This is not the case for advanced AI systems, at least presently. The speed of technological development, combined with the slowness of procedures in traditional SDOs, means they will not be capable of producing technical standards with sufficient detail to provide guidance on crucial aspects of governance, like cutting-edge model testing and evaluation methods. Furthermore, while other products are developed to meet certain specifications, AI may reveal some of its product-relevant properties only after deployment (e.g., depending on the context of application or on its capacity to interact with humans) (Zielke, 2020). Because of this, international standards alone will be insufficient for guiding regulatory compliance and supporting best practice.

It is thus helpful to distinguish technical standards, which provide guidance on rules and processes, and technical specifications, that prescribe more granular technical requirements.[6] For example, the business consortium MLCommons has developed performance and safety benchmarks (MLCommons, n.d.) that can be used as part of the model evaluation process laid out in a technical standard. This observation is consequential, as it indicates that standards-making efforts from SDOs – including some of the work being undertaken by CEN-CENELEC to explicate the AI Act – will need to be complemented by more granular technical work.

6.2. The role of SDOs

Three possible conclusions could be reached from the above discussion, which we assess in turn. First, that traditional SDOs should be entirely displaced by new standards organisations or business consortia who have more agile procedures. This is what happened for Internet standardisation, where traditional SDOs like ISO attempted to develop internet standards but had little success because of support for newer organisations, like the Internet Engineering Task Force (IETF) (Murphy & Yates, 2009). This argument has some worth. Yet, the types of standards being developed for the Internet were interconnectivity standards, while for advanced AI are physical externality standards. This distinction is important as there are weaker incentives for companies to collaborate to develop robust physical externality standards without external pressure. Traditional SDOs already act as regulatory intermediaries for governments and have established certification processes, suggesting they can still play an important incentivisation role.

[6] We use the term roughly in line with ISO's understanding of a "document that prescribes technical requirements to be fulfilled by a product, process or service" (ISO/IEC Guide 2:2004)

Second, it could be concluded that traditional SDOs should better utilise their existing toolkits, like PAS and fast-track procedures, to more agilely develop standards. Similarly, some SDOs, like ISO, already develop technical specifications for areas where technical consensus is not yet present, which could reasonably be proposed as a solution to the granularity problem. However, efforts to speed up the standards-making process could weaken the robust consensus procedures that characterise SDOs, while still producing standards and specifications far slower than newer institutions, like industry consortia and AI safety institutes (AISIs). Related arguments have been made that SDOs should fundamentally reform their functions to more agilely address issues raised by AI (Prifti & Fosch-Villaronga, 2024), yet this is extremely challenging in practice and could undermine the core competencies of these institutions.

The third conclusion – which we consider most viable – is that SDOs should focus on developing governance mechanisms that align with their institutional competencies and partner with other organisations to support greater harmonisation (Roberts et al., 2024). SDOs should not focus on "leading-edge" issues that guide innovation or target specific emergent risks. Instead, they should centre their efforts on standardising best practices at a higher-level, including through developing process or management standards which target more well-established problems and are more flexible to technological development. This move from "leading-edge" to "trailing-edge" has been a general trend in SDOs like ISO since the 1990s (Murphy & Yates, 2009), given the challenges these institutions have faced in responding to new technologies.

SDOs should also form stronger relationships with institutions that possess competencies better suited to developing granular and agile technical specifications. For example, several business consortia and AISIs are staffed by highly technically competent teams who are agilely developing methods for evaluating advanced AI. Rather than competing with these institutions, SDOs should attempt to utilise them as informal regulatory intermediaries who can explicate technical details of the high-level standards they produce (Abbott et al., 2017).

The current landscape of partnerships between SDOs and institutions creating technical specifications or standards-like documents is mixed. The business consortia MLCommons and the Partnership on AI have liaisons to ISO/IEC's AI standards committees, which can support institutional alignment. However, relationships with AISIs, who have been developing specifications for model evaluations, appear weak. Further work could be done to align these interests, given the complementary work they are undertaking.

6.3. Institutional reform

While it would be detrimental to attempt to fundamentally reform the institutional competencies and functions of traditional SDOs, work can still be done to ensure that they more inclusively and effectively develop international standards. Several light-touch changes to the participation model in leading SDOs could help support more equitable, inclusive, and impactful standards-making.

As some of our interviewees confirmed, standard-setting processes are still dominated by corporate and civil society actors from North America and Europe, with corporate actors continuing to outweigh civil society and academic participants (Auld et al., 2022). Several participants emphasised the need for small, focused, and continued grants for academia as well as participants from the Global

South, as there are currently few examples of such funding.[7] Improving the training offering for potential standards developers could also reduce the learning curve in meaningfully participating. SDOs already offer a wide range of courses, including general courses on how to draft international standards and specific regional training for validation and verification bodies.[8] But our interviewees rarely mentioned them as useful sources for learning institutional processes. Standards awareness workshops could also be targeted at specific stakeholder groups, tailored to specific expertise gaps in standardisation. Work by the AI Standards Hub and NIST to broaden engagement in the standards-making process could also be adopted as models to be used. For instance, NIST's "zero drafts" initiative (NIST, 2024) – designed to broaden participation and speed up standards development through preliminary, stakeholder-driven drafts – could progressively be more substantially integrated into SDO processes. Additionally, initiatives could be pioneered by SDOs to formalise this type of broader participation, for example, by tailoring recognition to stakeholder needs, such as crediting academic contributions similarly to citations.

7. Conclusion

The rapid commercialisation of advanced AI systems has introduced both transformative opportunities and complex governance challenges. International standards, developed by SDOs, can play an important governance role. The robust procedures of SDOs and their strong international reputations means these documents can be used to support compliance through explicating national-level regulations, which can in turn enable interoperability between different jurisdictions. Certification programmes associated with advanced AI standards can also guide companies towards best practices in the absence of regulatory requirements. These features suggest that international standards can play some role in governing advanced AI (Cihon, 2019).

Nonetheless, the findings of this paper demonstrate that international standards also have several deficiencies, which could limit their impact in practice. Adoption incentives, participation issues, the value-laden nature of AI, and the speed of technological development all pose challenges to the development and use of effective international standards for advanced AI. Based on this, we have argued that SDOs should focus on developing international standards related to their institutional competencies. Namely, their work will prove most impactful when it is focused on higher-level procedures and management practices, which are certifiable. To maximise their impact, SDOs should also cultivate and deepen partnerships with more agile business consortia and AISIs, who are developing granular technical specifications. By working with these institutions, rather than attempting to produce similar governance documents, SDOs can strengthen coordination in the international system and provide greater clarity to organisations seeking to govern AI.

These findings are consequential, as they highlight the breadth of institutions and governance tools that are needed to effectively address advanced AI. They also underline the importance of coordination between international institutions, which has received limited attention in global AI governance literature (Roberts et al., 2024).

[7] E.g., https://standict.eu/standicteu-2026-7th-open-call

[8] https://learning.iso.org/

Reference list

Abbott, K. W., and Snidal, D. (2001). International ‘standards’ and international governance. *Journal of European Public Policy*, *8*(3), 345–370. https://doi.org/10.1080/13501760110056013

Abbott, K. W., Levi-faur, D., & Snidal, D. (2017). Theorizing Regulatory Intermediaries: The RIT Model. *The ANNALS of the American Academy of Political and Social Science*, *670*(1), 14–35. https://doi.org/10.1177/0002716216688272

AI Standards Hub. (2023, January 12). *Event review: Towards transparent and explainable AI*. Retrieved from https://aistandardshub.org/event-review-towards-transparent-and-explainable-ai

Anthropic. (2025). *Anthropic achieves ISO 42001 certification for responsible AI*. https://www.anthropic.com/news/anthropic-achieves-iso-42001-certification-for-responsible-ai

Auld, G., Casovan ,Ashley, Clarke ,Amanda, and Faveri, B. (2022). Governing AI through ethical standards: Learning from the experiences of other private governance initiatives. *Journal of European Public Policy*, *29*(11), 1822–1844. https://doi.org/10.1080/13501763.2022.2099449

Bartley, T. (2022). Power and the Practice of Transnational Private Regulation. *New Political Economy*, *27*(2), 188–202. https://doi.org/10.1080/13563467.2021.1881471

Berti, L., Giorgi, F., & Kasneci, G. (2025). *Emergent Abilities in Large Language Models: A Survey* (No. arXiv:2503.05788). arXiv. https://doi.org/10.48550/arXiv.2503.05788

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (No. arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

Buhl, M., Hilton, B., Masterson, T., & Irving, G. (2025). *How can safety cases be used to help with frontier AI safety? | AISI Work*. AI Security Institute. https://www.aisi.gov.uk/work/how-can-safety-cases-be-used-to-help-with-frontier-ai-safety

Büthe, T., & Mattli, W. (2011). The New Global Rulers: The Privatization of Regulation in the World Economy. In *The New Global Rulers*. Princeton University Press. https://doi.org/10.1515/9781400838790

Cantero Gamito, M., & Marsden, C. T. (2024). Artificial intelligence co-regulation? The role of standards in the EU AI Act. *International Journal of Law and Information Technology*, *32*, eaae011. https://doi.org/10.1093/ijlit/eaae011

Charnovitz, S. (2005). International Standards and the WTO. *GW Law Faculty Publications & Other Works*. https://scholarship.law.gwu.edu/faculty_publications/394

Christmann, P., & Taylor, G. (2006). Firm self-regulation through international certifiable standards: Determinants of symbolic versus substantive implementation. *Journal of International Business Studies*, *37*(6), 863–878. https://doi.org/10.1057/palgrave.jibs.8400231

Cihon, P. (2019). *Standards for AI Governance:International Standards to Enable GlobalCoordination in AI Research & Development* [Technical Report]. Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Coglianese, C. (2023). *Standards and the Law* (SSRN Scholarly Paper No. 4452726). Social Science Research Network. https://papers.ssrn.com/abstract=4452726

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

David, P. A., & Steinmueller, W. E. (1994). Economics of compatibility standards and competition in telecommunication networks. *Information Economics and Policy*, *6*(3), 217–241. https://doi.org/10.1016/0167-6245(94)90003-5

Duffer, S., Singh, A., & Hallinan, P. (2024, November 25). *AWS achieves ISO/IEC 42001:2023 Artificial Intelligence Management System accredited certification | AWS Machine Learning Blog.* https://aws.amazon.com/blogs/machine-learning/aws-achieves-iso-iec-420012023-artificial-intelligence-management-system-accredited-certification/

Dworkin, S. L. (2012). Sample Size Policy for Qualitative Studies Using In-Depth Interviews. *Archives of Sexual Behavior*, *41*(6), 1319–1320. https://doi.org/10.1007/s10508-012-0016-6

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1696–1705). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1166

FMF. (2024, November 8). Issue Brief: Components of Frontier AI Safety Frameworks. *Frontier Model Forum*. https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/

FMF. (2025). *Frontier Capability Assessments—Frontier Model Forum.* https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023). *Bias and Fairness in Large Language Models: A Survey* (No. arXiv:2309.00770). arXiv. https://doi.org/10.48550/arXiv.2309.00770

Goldfarb-Tarrant, S., Marchant, R., Sanchez, R. M., Pandya, M., & Lopez, A. (2021). *Intrinsic Bias Metrics Do Not Correlate with Application Bias* (No. arXiv:2012.15859). arXiv. https://doi.org/10.48550/arXiv.2012.15859

Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). *Social Biases in NLP Models as Barriers for Persons with Disabilities* (No. arXiv:2005.00813). arXiv. https://doi.org/10.48550/arXiv.2005.00813

Infosys. (2024). *Infosys Receives ISO 42001:2023 Certification for Artificial Intelligence Management System*. https://www.infosys.com/newsroom/press-releases/2024/receives-iso-certification-ai-management-system.html

ITU. (2022). *ITU-T Y.3000-series – Artificial intelligence standardization roadmap*. https://demo.ifgict.org/wp-content/uploads/2024/08/ITU-Artificial-intelligence-standardization-roadmap.pdf

Kerwer, D. (2005). Rules that Many Use: Standards and Global Regulation. *Governance*, *18*(4), 611–632. https://doi.org/10.1111/j.1468-0491.2005.00294.x

Koremenos, B., Lipson, C., & Snidal, D. (2001). The Rational Design of International Institutions. *International Organization*, *55*(4), 761–799.

Korinek, A., & Vipra, J. (2025). Concentrating intelligence: Scaling and market structure in artificial intelligence*. *Economic Policy*, *40*(121), 225–256. https://doi.org/10.1093/epolic/eiae057

Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). Global Challenges in the Standardization of Ethics for Trustworthy AI. *Journal of ICT Standardization*, *8*(2), 123–150. https://doi.org/10.13052/jicts2245-800X.823

McDuff, D., Korjakow, T., Cambo, S., Benjamin, J. J., Lee, J., Jernite, Y., Ferrandis, C. M., Gokaslan, A., Tarkowski, A., Lindley, J., Cooper, A. F., & Contractor, D. (2024). *On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI* (No. arXiv:2402.05979). arXiv. https://doi.org/10.48550/arXiv.2402.05979

MLCommons. (n.d.). Benchmark Work | Benchmarks MLCommons. *MLCommons*. Retrieved 26 May 2025, from https://mlcommons.org/benchmarks/

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). *Auditing large language models: A three-layered approach* (No. arXiv:2302.08500). arXiv. https://doi.org/10.48550/arXiv.2302.08500

Murphy, C. N., & Yates, J. (2009). *The International Organization for Standardization (ISO): Global Governance Through Voluntary Consensus*. Routledge.

NIST. (2024). NIST's AI Standards "Zero Drafts" Pilot Project to Accelerate Standardization, Broaden Input. *NIST*. https://www.nist.gov/artificial-intelligence/ai-research/nists-ai-standards-zero-drafts-pilot-project-accelerate

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

Prifti, K., & Fosch-Villaronga, E. (2024). Towards experimental standardization for AI governance in the EU. *Computer Law & Security Review*, *52*, 105959. https://doi.org/10.1016/j.clsr.2024.105959

Randall Henning, C., and Pratt, T. (2023). Hierarchy and differentiation in international regime complexes: A theoretical framework for comparative research. *Review of International Political Economy*, *30*(6), 2178–2205. https://doi.org/10.1080/09692290.2023.2259424

Review into bias in algorithmic decision-making. (2020). GOV.UK. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. *International Affairs*, *100*(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

Roberts, H., Ziosi, M., & Cailean, O. (2023). *A Comparative Framework for AI Regulatory Policy.* CEIMIA. https://ceimia.org/wp-content/uploads/2023/02/Comparative-Framework-for-AI-Regulatory-Policy.pdf

Schneider, J. (2025). *Generative to Agentic AI: Survey, Conceptualization, and Challenges* (No. arXiv:2504.18875). arXiv. https://doi.org/10.48550/arXiv.2504.18875

Schuett, J. (2024). *From Principles to Rules: A Regulatory Approach for Frontier AI* (No. arXiv:2407.07300). arXiv. https://doi.org/10.48550/arXiv.2407.07300

Soler-Garrido, J., De Nigris, S., Bassani, E., Sanchez, I., Evas, T., André, A.-A., & Boulangé, T. (2024). *Harmonised Standards for the European AI Act.* JRC Publications Repository. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430

ten Oever, N., & Milan, S. (2022). The Making of International Communication Standards: Towards a Theory of Power in Standardization. *Journal of Standardisation*, *1*. https://doi.org/10.18757/jos.2022.6205

UK AISI. (2025). *RepliBench: Measuring autonomous replication capabilities in AI systems | AISI Work.* AI Security Institute. https://www.aisi.gov.uk/work/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

Vandenbergh, M. (2007). The New Wal-Mart Effect: The Role of Private Contracting in Global Governance. *UCLA Law Review*, *54*(4), 913.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.*, 7.

WTO. (1995). *WTO | Technical Barriers to Trade.* https://www.wto.org/english/tratop_e/tbt_e/tbt_e.htm

Yates, J., & Murphy, C. N. (2019). *Engineering Rules: Global Standard Setting since 1880.* JHU Press.

Zielke, T. (2020). *Is Artificial Intelligence Ready for Standardization?* n: Yilmaz, M., Niemann, J., Clarke, P., Messnarz, R. (eds) Systems, Software and Services Process Improvement. EuroSPI 2020. Communications in Computer and Information Science, vol 1251. Springer, Cham. https://doi.org/10.1007/978-3-030-56441-4_19

Acknowledgements: We would like to thank our interviewees for taking the time to speak with us about this project and for providing such helpful insights.

Funding: this research has been supported by ISO's 2023 Research Grant. Huw Roberts' research is supported by an ESRC DTP Studentship.