This is a preprint version of a chapter accepted for publication in the *Handbook of Human-Centered Artificial Intelligence* (Springer Nature, forthcoming 2026). Please cite the published version when available.

Strategies for Harmonizing Fragmented AI Ethics Frameworks, Standards, and Regulations

Daniel S. Schiff[1]

Purdue University, 2228 Beering Hall, West Lafayette, IN 47907

dschiff@purdue.edu

https://orcid.org/0000-0002-4376-7303

Abstract: AI governance is increasingly shaped by a patchwork of ethical frameworks, standards, and regulations, with overlapping demands from technical standards bodies, industry consortia, and governments. A key piece of this puzzle is *standardization*: as regulators increasingly delegate governance to standards development organizations (SDOs) in response to rapid AI innovation, hundreds of potentially overlapping standards proliferate, creating redundancy, organizational confusion, and superficial compliance, while audits and certifications struggle to meaningfully assess adherence. Drawing on an analysis of over 500 AI standards across multiple domains and issuing bodies, this chapter diagnoses five interrelated challenges complicating AI governance: the persistent difficulty of translating contested sociotechnical concepts like fairness, well-being, and transparency into actionable standards; the proliferation of voluntary frameworks with limited enforceability and vague operational guidance; decision paralysis as organizations confront competing and overlapping standards; standards development cycles that lag behind fast-moving AI innovation; and geopolitical competition that fragments international coordination. To address these challenges, the chapter offers a detailed set of institutional design strategies, grounded in current governance practice, for both harmonizing—and *humanizing*—AI governance. These include fostering deeper coordination among SDOs, supporting satisficing and layered framework adoption tailored to organizational capacity, establishing living standards and rapid-response taskforces to enable adaptive updates, strengthening auditing ecosystems through independent accreditation, transparency, and public reporting, and embedding stakeholder participation directly into governance workflows and organizational implementation processes. Crucially, AI governance must itself be human-centered, designed to be legible, adaptable, and actionable for the real humans responsible for implementing, auditing, and overseeing AI systems.

Keywords: AI governance, AI ethics, AI standards, standards development organizations, auditing and certification, human-centered AI, AI regulation, risk management, regulatory frameworks, responsible AI

[1] Transparency statement: This chapter was developed with assistance from AI tools, which were used to support literature review, refine ideas, draft content, and provide feedback. All outputs were critically reviewed.

1. Introduction[2]

In the 21st century, artificial intelligence (AI) has transitioned from a relatively niche field into a pervasive 'force' reshaping industry, public services, and everyday human life. Unsurprisingly, given the huge array of possible effects, we have witnessed an arguably unprecedented proliferation of ethical frameworks, regulatory initiatives, and—key to this chapter, standards. These governance tools are broadly aimed at mitigating AI's risks and maximizing its societal benefits in line with human-centered goals articulated by AI ethics and 'responsible' or 'trustworthy' AI discourse. In the coming years, the governance of AI systems and applications will become increasingly layered and fragmented, with individual use cases, systems, and organizations potentially subject to dozens of overlapping standards, frameworks, and regulatory requirements. These tools, ranging from informal to mandatory, include comprehensive frameworks such as the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF); subnational and industry-specific regulations; internal organizational governance policies and software-level ethical design standards; and emerging multinational legal regimes such as the European Union AI Act.[3]

What many of these efforts have in common is an effort to standardize best practices or requirements through de facto or explicit *standards*. While the definition and history of standards is complex (Brunsson, 2002), modern (20th century) standards typically refer to codified rules, established to enable consistent application of certain policies and practices. They constitute a form of social regulation that can "substitute for various other forms of authoritative rule" (Timmermans & Epstein, 2010, p. 71). Standards from technical bodies (i.e., "technical standards") often articulate specifications for the design of systems, their performance, their interoperability, and their safety, as well as organizational governance and risk management processes. They are often developed through consensus-based processes, including by industry or government consortia. And they may be binding or voluntary. Standards development organizations (SDOs) like the Institute of Electrical and Electronics Engineers (IEEE) and the International Organization for Standardization (ISO) have played an early and prominent role in shaping AI governance, and are likely to play an increasingly critical role (Narayanan et al., 2023).

It is also helpful to distinguish between process-based and outcome-based (or performance-based) standards. Process-based standards require organizations to follow specified procedures—such as conducting a human rights impact assessment, documenting data provenance, or engaging stakeholders—without necessarily ensuring specific outcomes. Outcome-based standards, by contrast, define performance targets or thresholds that systems must meet—like achieving certain safety levels or fairness metrics—while giving organizations latitude in how to comply. This affects compliance dynamics: outcome-based standards can drive contextual adaptation and creative innovation, but

[3] For an example of the cumulative complexity involved in governing a single technological tool, see Biddle et al. (2010), who identified over 250 distinct technical standards implicated in the design of a single laptop, and estimated that the total number of relevant standards is likely substantially higher.

may be hard to audit, susceptible to gaming, or leave organizations uncertain about implementation. Process-based standards, in comparison, often support clearer process verification, but risk rigid, superficial, or checkbox compliance if not well designed. For instance, the US NIST AI RMF exemplifies a process-based approach, whereas standards like ISO/IEC TR 24029-1 on robustness testing or ISO/IEC TR 24027 on bias assessment introduce performance-related expectations, though they often stop short of specifying precise numerical thresholds or binding outcome targets. Many real-world standards blend both modes, requiring robust organizational procedures and measurable system performance where feasible.

While the proximal focus of this chapter is on standards, this concept is treated broadly at times due to the complexity and overlap across different types of soft (i.e., voluntary or informal) and hard (i.e., formal or mandatory) policy instruments. Standards may be designed by governments, regulators, or industry consortia, may be voluntary or mandatory, and may cover both technical and process-based dimensions. They also interact with laws and regulations (which may reference, encourage, or require standards), as well as with organizational frameworks, ethical principles, and best practice guidance (which may lixewise incorporate or inform standards). These categories frequently blur, particularly in the AI governance space where instruments proliferate and hybridize. This chapter centers primarily on standards from technical bodies (SDOs)—including both technical system standards and organizational process standards. Yet it also necessarily addresses their interaction with regulatory frameworks, auditing and certification ecosystems, and organizational and ethical frameworks. To support conceptual clarity, Table 1 offers a simplified typology of key governance instruments relevant to AI, recognizing that these boundaries are often porous in practice.

Table 1. Typology of AI governance instruments and their roles

Instrument type	Typical source(s)	Binding nature	Primary function	Example(s)
Laws & regulations	Legislatures, regulatory agencies	Mandatory (hard law)	Define legal obligations; enable enforcement	EU AI Act; GDPR
Technical system standards	SDOs (ISO, IEC, IEEE, national standards bodies)	Voluntary; may be referenced in law (becoming de facto mandatory)	Define system specifications and testing (e.g., interoperability, safety, performance)	ISO/IEC TR 24029-1; IEEE P2801; ISO/IEC 25012
Organizational process standards	SDOs, public agencies (e.g., NIST), and multi-stakeholder initiatives / NGOs	Voluntary; may be referenced in law	Guide organizational processes (e.g., risk management, governance, auditing)	NIST AI RMF; ISO/IEC 42001; IEEE 7010
Certification schemes	SDOs, third-party auditors, regulators	Voluntary or required by regulation or market forces	Provide verification of conformance; signal trustworthiness; enable market discipline	IEEE CertifAIEd; EU AI Act conformity assessment
Ethical frameworks / principles	Governments, international organizations,	Voluntary (soft law / normative guidance)	Articulate high-level ethical values (e.g.,	OECD AI Principles; UNESCO AI

	industry consortia, NGOs, academic bodies		fairness, well-being, accountability)	Principles; IEEE Ethically Aligned Design
Organizational frameworks / internal standards	Private-sector organizations, professional bodies	Voluntary (internal governance)	Guide internal policies, practices, and culture; support signaling and compliance readiness	Microsoft Responsible AI Standard; Google AI Principles
Best practice guidelines / implementation tools	Regulators, NGOs, industry consortia, think tanks	Voluntary	Provide actionable guidance, operational tools, and implementation support for applying standards or principles	NIST AI RMF playbook; Model cards; Partnership on AI implementation tools

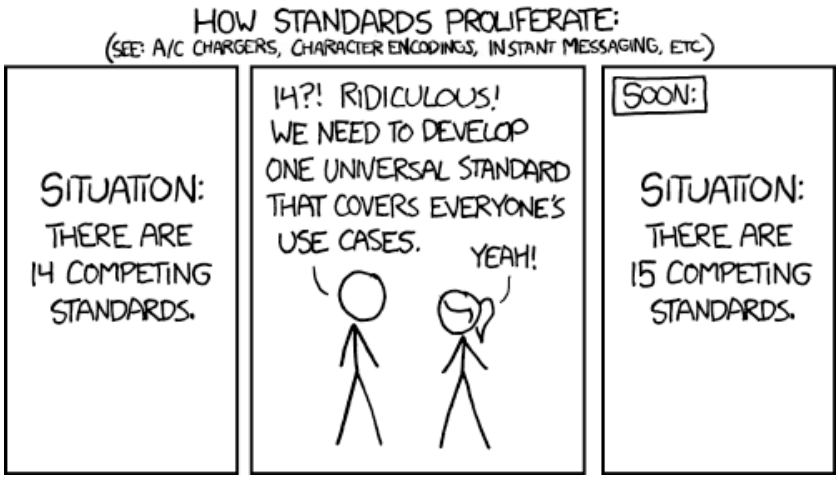
Note: The boundaries between these categories are porous in practice. Some actors (e.g., SDOs or regulators) may contribute to multiple types of instruments (e.g., both standards and associated implementation tools), and some instruments may shift across categories over time (e.g., voluntary frameworks becoming referenced in law).

At the heart of this rapidly-evolving landscape lies an interesting dilemma which, this chapter argues, is manifesting in a consequential shift in responsibility. Namely, policymakers, AI developers, and other stakeholders, recognizing that legislation can lag behind technological progress (G. E. Marchant et al., 2011), are relying heavily on technical bodies and industry consortia to translate broad consensus ethical principles for AI into workable standards. For instance, while stakeholders broadly agree that AI should be safe, fair, and transparent, the dedicated fields of effort aimed at addressing these issues may take decades to mature, with effort by thousands of individuals. Meanwhile, AI's penetration in society is already massive; governance cannot afford to wait (Bolte & van Wynsberghe, 2024). To address this gap, decision-makers are essentially "kick[ing] the can down the road" (Laux et al., 2024, p. 1) by asking other regulatory actors to develop best practices (Clouser McCann & Shipan, 2022). Yet this delegation shifts responsibility (Santoni de Sio & Mecacci, 2021) from publicly-accountable legislators to relatively opaque (and sometimes captured) technical bodies, a transfer that may or may not result in greater rigor, inclusivity, or genuine human-centered governance.

Consequently, SDOs and other standards-bearing efforts wield growing power in shaping AI governance, determining not just whether—but *how exactly*—principles like fairness, privacy, or accountability are instantiated in design practices. The burden on these organizations to get things right is immense (Stahl, 2023; Vallor & Ganesh, 2023). This delegation of responsibility also means that organizational leaders, developers, and auditors face a dizzying array of decisions. They must choose which standards or frameworks to adopt, how to align them with internal processes, and how to demonstrate compliance or certification to external stakeholders (Widder & Nafus, 2023). In short, a huge portion of decision-making around AI governance now depends on how successful SDOs and standardization efforts are at solving the problems of ethical, safe, and human-centered AI.

Extending from and complicating this issue is the plethora of frameworks, laws, and unsettled standards that an organizational actor must consider (Cihon et al., 2020; Corrêa et al., 2023; Jobin et al., 2019). Resultantly, "getting governance right" involves more than 'simply' adhering to 'the' established (single) standard; it rather requires calibrating multiple overlapping frameworks in a way that is both meaningful and practical for the humans who must implement them. Such considerations importantly move beyond their origin (in conceptual frameworks and high-level policies) into ostensibly concrete auditing and certification ecosystems (Blösser & Weihrauch, 2023; D. S. Schiff et al., 2024). These standardization efforts are thus becoming pivotal in operationalizing human-centered AI ethics and governance commitments (Xu, 2019; Xu et al., 2023).

Fig. 1. The problem with harmonizing standards



Note: Source: https://xkcd.com/927

However, auditing and certification efforts must grapple with overlapping standards that reference conflicting or cascading requirements, vague operational guidance for sociotechnical concepts like fairness or well-being, and fragmented certification schemes whose credibility and scope are uneven across jurisdictions and sectors (as explored in later sections). Done poorly, the scramble for compliance can devolve into a check-the-box exercise that overlooks ethical nuance (Delmas & Burbano, 2011; Kijewski et al., 2024), or that blows out into a huge array of poorly defined and bloated requirements (Bietti, 2020). Done well, however, standards and audits can form a cohesive system that not only guards against foreseeable harms but also upholds the dignity and agency of those affected by AI systems, in line with human-centered goals for AI (Shneiderman, 2020; White House, 2022).

In sum, and despite promising developments, the ecosystem remains fragmented and fraught with challenges. To help scholars, SDOs, and other governance actors reflect on these hurdles and contemplate their solutions, this chapter reviews five such challenges. First, many ethical concepts central to AI—fairness, human well-being, transparency—prove difficult to capture in purely technical terms, despite repeated and arguably naïve attempts. In trying to standardize these sociotechnical constructs, organizations risk glossing over political and social complexities and violating human-centricity. Emerging approaches are attempting to narrow this gap by enumerating sociotechnical

issues in greater detail and creating adaptive and scalable audit and evaluation frameworks grounded in real-world incidents, but it remains to be seen whether social and systemic issues can be technically or functionally embedded in this fashion to a sufficient degree. Second, most standards remain voluntary, raising persistent fears of "ethics washing," selective compliance, or reliance on untested or inadequate frameworks. Third, when standards proliferate without clear coordination, practitioners and organizations become overwhelmed by competing or duplicative requirements, and may be poorly positioned to identify which standards are high-quality or relevant. Fourth, the so-called pacing problem remains relevant even with efforts toward more adaptive modes of governance: as AI evolves rapidly, it outstrips the comparatively slow processes of consensus-building and standard revision. Finally, securing agreement among stakeholders—ranging from tech giants and civil society groups to regulators scattered across diverse geopolitical contexts—often proves elusive, complicating efforts to develop truly harmonized or human-centered global frameworks.

This chapter not only offers a diagnosis of these structural challenges; it also proposes a set of grounded strategies to address them. I*t argues that our approach to designing, selecting, and implementing standards (and associated frameworks and regulations) must itself be human-centered*, extending previous arguments about human-centered AI (Capel & Brereton, 2023; Xu et al., 2023). Section 2 begins by tracing the causes and consequences of the proliferation of AI ethics standards and regulations, spotlighting how delegated governance and competitive standard-setting threaten redundancy and confusion. Section 3 reviews the aforementioned challenges that stem from translating complex ethical values into actionable requirements, negotiating voluntary compliance, keeping pace with technological changes, and securing consensus.

Next, Section 4 proposes a suite of strategies for advancing a more harmonized and genuinely human-centered approach to AI standards and governance. Human-centered governance here refers not merely to protecting end users through the articulation of ethical principles or goals (i.e., *human-centered AI*), but rather to structuring governance processes, auditing mechanisms, and organizational adoption pathways that are intelligible, practical, and responsive for the diverse human actors tasked with implementing them. The chapter outlines solutions such as strengthening international coordination and crosswalks across frameworks, adopting satisficing and layered approaches to organizational framework selection, establishing adaptive mechanisms like living documents and rapid-response taskforces to address the pacing problem, and enhancing the auditing ecosystem through independent oversight, transparency, and professional development. By embedding these human-centered strategies *into the design of governance itself*, the chapter argues that institutions can reduce fragmentation, avoid paralysis amidst proliferating standards, and build more durable and accountable systems capable of safeguarding human well-being in the face of rapid AI innovation and governance experimentation.

2. The Proliferation of AI Ethics Standards, Frameworks, and Regulations: Causes and Consequences

The rapid evolution of AI in the 2020s, bolstered by breakthroughs in deep learning, the data-rich digital ecosystems of the 2010s, and by accelerated funding, has catalyzed a rush to articulate ethical guidelines, standards, and regulations. Policymakers, eager to capitalize on AI's benefits while ostensibly alarmed by the possibility of AI-driven harms, have sought to intervene before risks intensify or competitors take leadership. Meanwhile industry actors, eager to forestall heavy-handed rules or preserve competitive advantage, have likewise raced to shape self-regulatory frameworks that appear robust yet remain business-friendly (G. Marchant, 2019; D. S. Schiff, 2023). Beyond an AI arms race itself, these forces have thus produced what we might term a regulatory arms race, wherein national governments, multilateral organizations, industry consortia, academics, and professional bodies each attempt to lead the creation of the definitive rulebook for safe and responsible AI. The results are manifold, from the European Union's landmark AI Act and the US NIST AI RMF to specialized technical system standards and organizational process standards driven by SDOs (Arnold et al., 2024; Morandín-Ahuerma, 2023).

Essentially, in light of legislative lag and the technical complexity of AI, policymakers in many jurisdictions have begun to delegate intricate governance decisions to specialized organizations (Laux et al., 2024). This is a consistent feature of the modern state, where policymakers with minimal time, expertise, and attention increasingly rely on specialist bureaucrats or industry experts to 'figure out the details.'[4] As such, rather than codify every nuance of AI safety, fairness, or transparency directly in law, an unrealistic and inefficient approach, these policymakers perhaps justifiably rely on SDOs (and, increasingly, private companies given their heightened access and expertise in some domains (Anderljung et al., 2023)) to devise standards that operationalize broad principles. Through consultation and consensus-building processes, groups such as IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2019) or ISO/IEC JTC 1/SC 42 (Dudley, 2024) are producing standards that purport to translate human-centered norms into requirements, and that seek to become globally-relevant standards.

Another central factor fueling this proliferation is the incentive for various entities to assert leadership in AI governance (D. S. Schiff et al., 2022; Seo & Koek, 2012). Governments wishing to be seen as global technology hubs often champion national or regional regulations, aiming to export their legal and ethical values worldwide or to present themselves as global innovators (Schirm, 2010) even if they are small companies or countries. At the same time, private-sector coalitions and SDOs recognize that whoever sets the de facto technical benchmarks exerts considerable power over the direction of AI governance and consequently even AI innovation itself. This means that different actors, all purporting to serve the public interest, may effectively be competing to define the boundaries of responsible AI, and not always in the public interest (Cihon et al., 2021; von Ingersleben, 2023).

Organizations involved in these standards-setting efforts operate at multiple levels. At the international level, major standards bodies such as the International Organization for Standardization (ISO), the International Electrotechnical

[4] For instance, an analysis of major US laws from 1947 to 2016 finds that more than 99% delegate meaningful responsibility from legislators to agencies (Clouser McCann & Shipan, 2022).

Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), and the International Telecommunication Union (ITU) develop global guidelines for technology, safety, interoperability, and AI governance. At the regional level, a variety of standards development organizations also coordinate efforts cross-nationally. For example, in Europe alone, bodies such as the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI) play particularly influential roles given the EU's leadership in AI regulation. Meanwhile, industry-specific bodies, such as ASTM International and SAE International, focus on technical standards for particular sectors including materials, healthcare, and transportation. Finally, at the national level, organizations like the American National Standards Institute (ANSI, United States), the British Standards Institution (BSI, United Kingdom), the Deutsches Institut für Normung (DIN, Germany), the Japanese Industrial Standards Committee (JISC, Japan), the Standardization Administration of China (SAC), and the Korean Standards Association (KSA) develop country-specific standards that ideally seek alignment with international frameworks.

As captured by groups like OCEANIS (Open Community for Ethics in Autonomous and Intelligent Systems, a global forum for primarily SDOs sharing standards-related information and promoting cooperation) and the similarly-oriented AI Standards Hub operated by the Alan Turing Institute, AI standards might differ by domain, application, scope, topic, stage of development, issuing body committee, and type of standard (including voluntary or mandatory nature), as well as along other dimensions. For instance, the AI Standards Hub alone identifies more than 500 relevant standards (Alan Turing Institute, 2024). This includes, for example, standards[5] focused on:

- Applications such as speech recognition, facial recognition, video analytics, or robotics;
- Sectors such as education, energy, financial services, healthcare, manufacturing, transportation;
- Scopes such as corporate governance, cybersecurity, IT governance, project management, risk management, and software;
- Topics such as accountability, procurement, bias, data processing, documentation, robustness, safety, skills, and software architecture; and
- Emphases such as terminology, architecture, measurement, process, and performance[6]

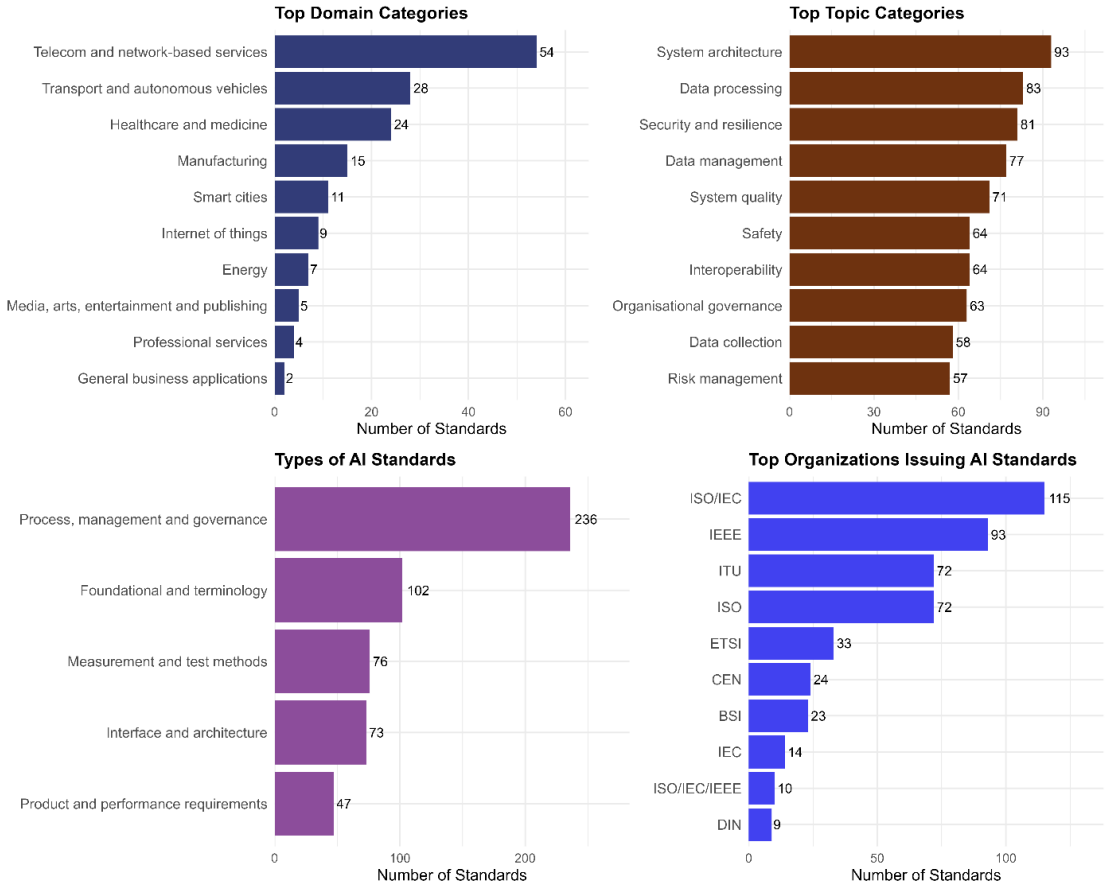
And while this section focuses primarily on standards, it is important to recognize that AI governance more broadly encompasses a range of related (and equally diverse) instruments—including laws, certifications, ethical principles, and best practice tools—that interact with and shape the evolving standards landscape (as reflected in Table 1).

[5] For instance, BS 30440 (Validation framework for the use of AI within healthcare), CEN/CLC/JTC 21 N 148 (AI-enhanced nudging), PAS 1881 (Assuring the operational safety of automated vehicles), and IEEE P3462 (Recommended Practice for Using Safety by Design in Generative Models to Prioritize Child Safety) are amongst more than 500 standards summarized in Figure 2. Both OCEANIS and the AI Standards Hub retain repositories; the latter database is more robust and therefore used for analysis here.

[6] The AI Standards Hub provides definitions of most of these categories in its taxonomy, though admittedly there is both practical and conceptual overlap in the taxonomy—and in the standards themselves—which makes precise delineation difficult.

To help illustrate some of this complexity, Figure 2 analyzes more than 500 standards[7] captured by the AI Standards Hub. Amongst the most active SDOs are ISO/IEC, IEEE, and ITU (Figure 2, bottom right), each with more than 50 related standards efforts. These efforts span a wide array of domains and sectors, including telecom, autonomous vehicles, healthcare, and manufacturing (Figure 1, top left); address diverse governance and technical topics such as system architecture, data management, safety, and organizational governance (Figure 2, top right); and encompass different types of standards, from process and management approaches to foundational terminology and technical test methods (Figure 2, bottom left). Together, these patterns further illustrate the breadth, heterogeneity, and complexity of the emerging AI standards ecosystem.

Fig 2. Distribution of 500+ AI Standards by Domain, Topic, Type, and Issuing Body



Notes: Data source and taxonomy: AI Standards Hub; analysis by the chapter author. This figure summarizes the heterogeneity of AI-related standards captured by the AI Standards Hub across its taxonomy of domains, topics, standard types, and issuing bodies. Categories are limited to the top 10 for visibility (or top 5 in the bottom-left panel).

[7] Note that the AI Standards Hub also retains different versions of some standards (e.g., IEEE 3302 V1, V1.1, and V1.2) as separate entries, so the number of *unique* standards is lower.

Counts do not sum to the full number of standards analyzed (n = 517) due to missing data (e.g., pre-draft standards), and some standards analyzed are successors of prior versions, so the number of unique standards is smaller; Figure 1 should be interpreted accordingly.

Some degree of diversity in standards is to be expected given the scope of AI applications and governance contexts, and this pluralism can even be beneficial in promoting flexibility and sector-specific relevance. Yet not all such diversity is benign or sound. In many cases, overlapping or duplicative standards may reflect uncoordinated processes, fragmented governance efforts, or strategic behavior by stakeholders seeking to shape norms in their favor. As a result, this profusion of guidance can leave practitioners and policymakers alike unsure about which standards are authoritative and of high quality, how they interrelate, and how to navigate an increasingly crowded landscape.

That is, in principle, this explosion of regulatory documentation reflects a genuine desire to steer AI toward constructive social outcomes, with thoughtful, human-centered governance across numerous important domains (D. S. Schiff et al., 2022). *In practice*, it is likely to generate confusion for organizations trying to comply with a patchwork of overlapping and sometimes contradictory directives. At the heart of this confusion is the growing burden placed on corporate compliance teams, developers, and auditors as they attempt to navigate, prioritize, and reconcile proliferating standards.[8] They must decide which standards are relevant and aligned with regulations; which to prioritize; how to adapt them to specific organizational, sectoral, and application contexts; and how best to integrate additional frameworks addressing domains such as data governance, cybersecurity, or environmental impact (Park, 2024). As a consequence, while this diffusion and delegation of authority can foster specialization and nimbleness, it also fragments responsibility (Santoni de Sio & Mecacci, 2021). SDOs are working to translate broad ethical principles and regulatory guidance into codified requirements, but with limited coordination across jurisdictions, sectors, and applications, overlapping and sometimes inconsistent standards accumulate. As a result, organizations, compliance teams, developers, and auditors, must navigate and operationalize ambiguous and competing frameworks, a process that risks sacrificing rigor and the very human-centered values these standards aim to promote.

Complicating matters is the growing importance of auditing and certification ecosystems that link AI governance commitments to practical enforcement (D. S. Schiff et al., 2024). Even if a given standard is nominally voluntary, an external auditing firm may measure a product's alignment to that standard, effectively imposing market-based discipline. Organizations that earn a "trustworthy AI" seal for themselves or their products, for instance, can leverage that accreditation to reassure customers and investors (Wittenberg et al., 2024). Yet with multiple frameworks on offer, it remains unclear which certifications carry genuine weight or whether some might devolve into superficial box-ticking (Raji, 2022). In addition, certifiers themselves must decide how to interpret codes that can be conceptually vague: terms like "well-being" or "algorithmic justice" do not map neatly onto universal sector- or technology-agnostic checklists. Under these conditions, a robust auditing ecosystem has tremendous promise for mainstreaming

[8] Not to mention the burden placed on policymakers, scholars, civil society watchdogs, or the general public.

AI ethics but risks further fragmentation if different auditing bodies champion divergent standards (von Ingersleben, 2023). Arguably, incentives from companies and even auditors in the emerging auditing ecosystem could tend towards lower and simpler standards, though such an approach would be shortsighted, failing to produce meaningful signals that differentiate high and low performers.

On one hand then, various actors may jockey for authority in a constructive sense, progressively filling gaps left by slowly evolving legal frameworks. On the other hand, many of the resulting guidelines and processes may generate as much confusion as clarity. Their value depends on how effectively they ultimately align with actual human needs and capacities. A major question, then, is *how to approach these proliferating frameworks in a manner that upholds human-centered values* without overburdening the very people who must implement or benefit from these systems. Organizations and developers—especially those working in smaller companies or regions with fewer resources—can find themselves overwhelmed by the volume and complexity of guidance. End users and affected communities, particularly those that are marginalized or less technically informed, may have difficulty tracking what any given standard actually promises (Bogina et al., 2021; Liesenfeld & Dingemanse, 2024) with regard to fairness or data protection, much less evaluating whether a framework is effective or AI product ethically sound.

In short, AI governance structures need to be legible, coherent, and navigable for all stakeholders if they are to realize the ethical aspirations that motivated them in the first place. As this chapter contends, the fragmentation problem is not just a technical matter of integrating multiple standards; it is fundamentally a human issue, requiring a governance approach designed to be intelligible and workable for a diverse array of individuals, organizations, and social contexts. The next section details how some of the aforementioned challenges manifest in practice and explores why establishing genuinely human-centered governance is an ongoing and urgent project.

3. Challenges of the AI Standards and Regulatory Ecosystem

In the midst of the prolific rush toward more comprehensive AI standards and regulations, several persistent challenges stand in the way of creating a coherent governance environment. The issues explored here—ranging from the technical translation of ethical ideals to geopolitical tensions—(begin to) reveal how difficult it is to align divergent stakeholder interests in a potentially ever-accelerating technological landscape. Appreciating these complexities is pivotal for anyone seeking to build a human-centered system of AI standards, as these hurdles are likely to persist unless (and even if) they are addressed head-on. In the worst case, efforts to improve AI standards, regulations, and governance can backfire. One attempt of this chapter is to lower the likelihood of this unfortunate irony.

3.1 Translating Sociotechnical Issues into Technical Standards

One of the most odd and vexing barriers in AI governance is the ambition to translate often contested ethical concepts—fairness, well-being, transparency, justice—into concrete technical requirements. Although numerous proposed guidelines often *promise* clarity, attempts to standardize what are, at root, sociotechnical constructs invariably risk oversimplification. Here, sociotechnical refers to the unavoidable interdependence (Cooper & Foster, 1971) between social systems (e.g., people, organizations, institutions, policies, human goals) and technical ones like AI (Selbst et al., 2019), specifically with respect to their mutually constitutive origins, characteristics, and outcomes.

The first element of this translation effort is the *inherent complexity and political contestatio*n involved in ethical concepts. Bias, for instance, can signify anything from statistical disparities in false positive rates to structural inequities rooted in historical discrimination (Agarwal & Agarwal, 2023; Mehrabi et al., 2021).[9] Transparency may entail clarifying the data provenance for a machine learning system, providing public-facing audit reports, offering detailed model explanations, or simply letting a user know that an AI system is being used (Hacker & Passoth, 2022; Nannini et al., 2023). Each of these notions is shaped by cultural context, stakeholder values, and political agendas (Brauner et al., 2024). Consequently, no single "objectively correct" criterion can capture the full complexity of bias or transparency, and even an extensive quantitative suite of such criteria is unlikely to suffice. When SDOs attempt to codify such concepts, they therefore face the dilemma of either issuing high-level, abstract guidance that is too vague to implement or else enumerating checklists that smooth over critical context and nuance.

This dilemma points to a problem that the AI ethics, safety, and policy communities have not adequately taken onboard despite ample debate—*a persistent and aspirational tendency to overestimate the capacity of technical standardization to formalize sociotechnical concepts* whose ethical complexity resists reduction to codifiable metrics or universal procedures. This dynamic is reinforced by engineers and organizational compliance teams who often push for unambiguous, testable benchmarks (Martínez-Plumed et al., 2021; Reuel et al., 2024). While this is an understandable goal, this pragmatic orientation risks obscuring subjective, contested, yet indispensable ethical dimensions central to AI governance. Some fairness metrics, for example, may be easy to measure in controlled settings, but typically fail to capture broader structural injustices (Selbst et al., 2019; Weinberg, 2022). Similarly, while safety metrics may seem straightforward in controlled settings, real-world deployments introduce complex and evolving interactions among AI systems, humans, and environments that make uniform safety standards difficult to define or sustain (Amodei et al., 2016; Park, 2024). Ironically, this very challenge was recognized in the original formulation of sociotechnical systems theory by Emery & Trist (1960): technical systems can only function effectively when jointly optimized alongside the social and organizational environments they shape and depend on.

Some emerging proposals and frameworks are seeking to close these kinds of gaps as more actors pay attention to societal and systemic risks. This includes efforts to better define sociotechnical risks and harms, to measure them, and to use those evaluations to inform practice. For instance, from a definitional perspective, the AI Risk Repository

[9] Mehrabi et al. (2021) offer 23 forms of bias to consider, and Agarwal & Agarwal (2023) proposed a 'seven-layer' model to address this topic alone.

(Slattery et al., 2025) captures more than 1,600 risks across seven domains and 24 subdomains based on 65 existing frameworks, with risks mapped to existing databases like the AI Incidents Database. Meanwhile, the AILuminate benchmark suite evaluates a set of 12 types of hazards, including "non-physical" and "contextual" hazards that might emerge from interaction with AI chat models (Ghosh et al., 2025). To operationalize these kinds of efforts, the concept of Evaluation Authorities has been proposed, wherein independent auditors would use incidents to create programmatic tests whose results would feed directly back into a company's AI system development (Chadda et al., 2024). These efforts do represent an increasingly serious effort to ground sociotechnical issues in contextually appropriate and feasible 'technical' processes (construed broadly), yet even the Herculean nature of the initial effort required—e.g., measuring 1,600 risks or synthesizing 65 framework—reveals the daunting barriers to formalization. It also raises another complication: should standards and metrics be personalized and centered on *individual* humans in their unique contexts ('human-centered')? Or should they be broad, sweeping standards that are attuned to a *collective* level ('humanity-centered')?

As such, even if highly sophisticated suites of quantitatively-oriented risks, harms, benchmarks, thresholds, and technical standards are developed, it remains questionable whether the collective result will be feasible or sufficient. The danger is that codification—especially in quantitative form—risks collapsing the moral seriousness of human-centered AI into procedural checklists, even when packaged as 'comprehensive' frameworks or narrative-driven guidelines. However expansive or well-branded, such approaches may ultimately amount to little more than formalized box-ticking. Ultimately, AI systems and organizations might pass these kinds of auditing tests yet fail to protect actual communities and societies (Madaio et al., 2024). This danger is especially acute if the most thoughtful and ambitious toolkits envisioned by forward-thinking proponents fail to achieve widespread adoption, while simpler checklists and the gameable benchmarks thrive. In essence, technical codification may be fundamentally an insufficient tool, and one that is susceptible to capture (Martínez-Plumed et al., 2021).

Moreover, under these conditions of increased codification, issues of contested and ambiguous values remain unresolved. As Matus and Veale (2021, p. 1) note in their review of lessons for AI governance drawn from sustainability discourse, "machine learning is characterized by difficult or impossible to observe credence properties, such as the characteristics of data collection, or carbon emissions from model training, as well as value chain issues, such as emerging core-periphery inequalities, networks of labor, and fragmented and modular value creation." This observation illustrates the broad and complex ethical dimensions implicated even within a single domain of impact. Further complicating matters is the implicit requirement that SDOs or other policymakers must make determinations about which values take precedence and which interpretations of those values are correct (Laux et al., 2024), raising further questions about when it might be more appropriate to draw on public preferences, expert guidance, or some negotiated consensus (Bogucka et al., 2024).

It is perhaps unsurprising, then, that numerous standards and regulatory frameworks have adopted sociotechnical and process-based approaches (T. Goodman, 2023; Stranieri & Sun, 2022) in lieu of performance-based ones. The first

international AI ethics standard, the IEEE 7010-2020 standard, applies a sociotechnical framework (D. Schiff et al., 2020), as does the prominent NIST AI RMF (NIST, 2023). The EU AI Act likewise requires human rights assessments which lend themselves to qualitative evaluation. Yet these more holistic approaches unfortunately introduce an inverse problem relative to technical standardization: they often yield requirements that are insufficiently defined and difficult to measure, a concern frequently raised in critiques of human-centered AI and AI ethics principles. Sociotechnical and process-based standards may be perceived as even more subjective, more easily diluted or gamed, and harder to compare across organizations or systems (E. P. Goodman & Trehu, 2022). Alas, there is no one-size-fits-all approach for sociotechnical governance either.

Nevertheless, one key takeaway is that human-centered AI cannot rely on technical metrics alone. Attempts to (merely) quantify the complex dimensions of ethics and AI's social implications risk flattening entire disciplines of inquiry and bypassing centuries of accumulated human experience. Instead, a more workable middle ground may lie in process-based approaches: for instance, human rights impact assessments or well-being assessments can be administered through formalized steps, even using structured tools such as checklists, while still allowing necessary qualitative flexibility for contextual interpretation. A pragmatic approach, then, is to quantify where quantification is appropriate (e.g., allowable CO_2 emissions), while relying on process-based or sociotechnical frameworks where qualitative evaluation is more suitable (e.g., environmental justice assessments). Yet while recognizing for which domains quantification or technical standards are appropriate—or not—may be a plausible first step, the contours of these determinations remain complex and are left to future research.

3.2 The Voluntary Nature of Many AI Standards

A second, closely related challenge is that most AI standards remain voluntary. This is partly by design, as SDOs typically rely on 'consensus' procedures to accommodate a broad range of industry stakeholders who are wary of rigid requirements that might stifle innovation. But this voluntariness can undermine accountability and, ironically but unsurprisingly, even encourage a form of ethical shirking or minimalism.

First, when a standard is optional, *compliance can easily turn into box-ticking, public relations gestures, or superficial audits*, typically associated with "ethics washing" (E. P. Goodman & Trehu, 2022) A company can assert alignment with a certain ethical framework while doing little to mitigate tangible harms—arguing that they've "adhered to standards" even if said standard is insufficiently robust or its implementation inadequately verified. Such ethics-washing practices dilute accountability and risk eroding trust in the standards ecosystem among civil society, regulators, and end users. This makes it harder for genuinely robust standards and organizations who take human-centered AI seriously to distinguish themselves (Bietti, 2020; de Laat, 2021; Ibáñez & Olmeda, 2021).

Definitionally, *voluntary standards lack robust enforcement mechanisms*. Because they carry neither legal force nor (directly) financial or other penalties, organizations have ample freedom to cherry-pick particular frameworks or even convenient elements of a framework without genuinely changing their practices. Informal mechanisms like reputation,

customer trust, or fears of employee pushback are legitimate levers of action (D. S. Schiff et al., 2025) but often less impactful than command-and-control regulation or the threat of serious fines or lawsuits. Even well-intentioned organizations might struggle to implement ambitious guidelines without clear enforcement pathways or incentives. For instance, the challenges facing effective coordination between an internal ethics team, risk management team, and product development team could lead to lackluster attempts despite the prosocial aspirations of the majority of individuals. Take a commonly recommended strategy like drawing on public participation in AI development (Ouchchy et al., 2020). An ethics team might want to establish a diverse deliberative democratic body composed of members of the general public to advise on AI product risks or harms, while the risk team is unable to address issues around recruitment, compensation, or sharing of trade secrets. Meanwhile, the product development team might be sorely lacking in formal infrastructure and expertise in facilitating participatory processes. Without clear guidance and incentives, the aspiration for a participatory approach may remain notional.

To counter the downsides of voluntarism, different approaches have been offered. These include formalizing collective governance at the organizational level in lieu of government accountability, with increasingly binding requirements (J. B. Biddle et al., 2025), mandating the independence of auditors even for voluntary frameworks (D. S. Schiff et al., 2024), and holding benchmarks closely to avoid data contamination and overfitting or gaming of benchmarks (Chadda et al., 2024; Martínez-Plumed et al., 2021; Reuel et al., 2024). Yet over time, the ongoing proliferation, selective promotion, and ad hoc application of standards risks undermining the credibility of the AI standards ecosystem altogether if meaningful implementation and enforcement mechanisms are not established (Radclyffe et al., 2023). This problem becomes particularly acute when multiple standards exist in parallel, each claiming to represent "best practices" while leaving the hard work of operationalization to organizational discretion. In contrast, clear standards with serious penalties can make apparent that organizations need to hire, train, and reorganize in order to make safety, ethics, and audit procedures actionable (Armour et al., 2020).

Finally, voluntary approaches (alone) *may also hamper accountability and collective learning* in AI governance. Without clear enforcement or auditing mechanisms, it becomes difficult to pinpoint which standards prove most effective at reducing bias, enhancing transparency, or preventing harm. And because noncompliance rarely triggers formal penalties, the feedback loop that might spur standards revision or more rigorous oversight remains weak. As a result, monitors, be they from the government, civil society, or the media, are partially reliant on luck to even identify harms that occur. For instance, incident databases compiled by unofficial actors (Paeth et al., 2024) offer crucial visibility into emerging harms, yet without formal regulatory scrutiny, these monitoring efforts cannot fully substitute for systematic oversight and government-mandated enforcement.

3.3 Overlapping and Competing Standards

As discussed earlier, beyond issues of translation and voluntary adoption, *the sheer quantity of available AI standards and other governance tools introduces its own layer of confusion* (Birkstedt et al., 2023; Hernandez et al., 2024). Multinational corporations, for example, may face a dense forest of "best practice" frameworks—some broad, some

domain-specific, some emerging from non-governmental bodies, others tied to national or regional legislation. As depicted in Figure 1, there are also different AI-related standards (and regulations and frameworks) for different sectors such as energy (Pelekis et al., 2024) and education (Tong et al., 2024), and healthcare (Solanki et al., 2022; Tam et al., 2024). As organizations attempt to identify and adopt applicable standards for each relevant domain, sector, and regional context, they face an increasingly complex array of trade-offs (Oesterling et al., 2024). And this difficulty is compounded by the broader fragmentation of the AI governance ecosystem. Organizations must navigate not only multiple standards, but also a layered mix of certification schemes, ethical frameworks, regulatory requirements, and other governance instruments (see Table 1), each varying in scope, authority, and practical implications.

From corporate compliance teams to small engineering start-ups, practitioners must ultimately decide which set of requirements to prioritize, leading to confusion among practitioners as well as potential decision paralysis (Howard, 2020). These decisions often reflect a complex interplay of market positioning, regulatory context, and internal organizational priorities. For example, certification under a particular standard may signal credibility or confer market advantage in Europe but carry little weight in East Asia. Global standards from bodies such as ISO or IEEE may conflict with local regulatory requirements or fail to address region-specific legal or cultural concerns. Even within organizations, different teams may prioritize standards differently: legal teams may focus on cautious regulatory alignment, while data science teams emphasize technical benchmarks or performance metrics.. For time-strapped organizations, rational decision-making about which standard to follow becomes fraught when they are bombarded by multiple overlapping guidelines. This may result in organizations complying with the simplest or most binding standard (in a regulatory sense), or simply the most popular one. However, this does not necessarily mean they will adopt the best or most human-centered set of standards. Indeed, these decision heuristics may trend towards ease rather than quality (Gigerenzer et al., 2022).

As just one example, there are more than a dozen IEEE P7000-series projects alone, all of which emerged from the IEEE's Global Initiative on the Ethics of Autonomous and Intelligent Systems (Chatila & Havens, 2019), and aim to codify ethical practices in AI. Despite the foresight and ambition of this initiative, standards in this series—which focus on topics like bias, transparency, ethical design, well-being, privacy, and more—at times compete for the same conceptual space. For practitioners, it is not clear whether they should adhere to IEEE 7000, 7010, the entire 7000 series, or pursue certification through IEEE's own CertifAIed program, which trains assessors and offers accredited assessments but operates as a distinct certification framework. Nor is there a tool provided to help organizations pick between the 7000 series standards. Finally, each of these standards includes a list of normative and informative references, which point to *additional* standards that organizations should (informative) or must (normative) adhere to during compliance. These referenced standards often include yet more normative and informative references, creating a cascading "rabbit hole" of requirements. It is likely that many organizations do not adequately follow this complex web of referenced standards.

The proliferation of standards can also contribute to, and emerge from, fragmentation within organizations. Organizational units focused on privacy, ethics, risk management, or legal compliance may each gravitate toward particular standards that align with their distinct mandates. As a result, they may independently adopt different frameworks, or interpret and adapt the same framework in divergent ways, leading to variations in scope, tone, and evaluative emphasis even when addressing ostensibly the same concerns such as transparency, data governance, or bias mitigation. For instance, risk management teams may adopt regulatory-oriented, process-based frameworks like the NIST AI RMF, ethics professionals might emphasize holistic approaches such as IEEE 7010, and data scientists may focus on narrow technical testing standards for issues like algorithmic bias. These divergent approaches often resist integration, leaving organizations with fragmented compliance regimes that fall short of coherent governance.

3.4 The Pacing Problem and Difficulty in Updating Standards

Even if standards were perfectly aligned, which is far from the case, they would still confront a fundamental structural challenge: the remarkable speed of AI innovation. Model architectures, training paradigms, deployment strategies, and application domains continue to evolve so rapidly that conventional standard-setting processes, which often require years for consensus, risk near-immediate obsolescence. Terms like generative AI, frontier AI, general-purpose AI, etc., were almost entirely absent from the public radar until they quickly became state-of-the-art terms and the sine qua non of AI discourse. In essence, *AI risks advancing faster than standards can keep up*. In principle, governance bodies might anticipate upcoming trajectories of AI research, but uncertain or fast-shifting directions complicate such foresight, meaning that reactive vs. proactive standardization is more likely.

That is, while corporate research labs iterate on new architectures and deployment models within mere months, SDOs rely on cycles of proposals, reviews, and balloting (voting) that often span multiple years. This slow consensus-building process is not incidental; it reflects deliberate efforts to foster legitimacy and quality through broad stakeholder consultation and negotiation. Unsurprisingly, reconciling the divergent interests and views of large corporations, small startups, government agencies, civil society advocates, and academic experts cannot happen overnight. Yet the rapid cadence of AI's iteration likewise increasingly demands agile governance (Tong et al., 2024), creating a persistent tension between procedural legitimacy and practical responsiveness. Notably, it remains unclear which actors are best positioned to support this agility: while corporate experts may move quickly, they face conflicts of interest; academic experts may have fewer conflicts, but often operate at slower institutional tempos

As a result, by the time an AI ethics standard is finalized and published, the technological frontier may have shifted, introducing novel architectures, deployment contexts, or emergent risks that render earlier standards and guidelines incomplete or even counterproductive. This "pacing problem" traps SDOs into a reactive mode, continually playing catch-up with industry innovations and public controversies. The lag poses serious challenges for both organizations and the broader standards ecosystem. Most immediately, it discourages adoption (or even consideration) of now-outdated standards that may fail to address emerging risks. It also complicates harmonization efforts and exacerbates governance fragmentation, as multiple standards bodies pursue overlapping initiatives on misaligned timelines, often

unable to coordinate due to secrecy, competition, or information-sharing barriers during development. Perhaps most concerning, organizations and experts may forgo participation in standards development altogether, anticipating slow, uncertain processes with little return on their investment of time and resources.

These shortfalls intensify the necessity for flexible, "living" standards and highlight why any effort to harmonize frameworks must address updating mechanisms. As Paskov et al. observe, (2024, p. 2), familiar goals of standards like achieving "internal validity, external validity, reproducibility, and portability" may be insufficient on their own. Instead, "standards may best exist within an adaptive framework that can evolve with changes in technology, environment, and risk." At its core, then, the pacing problem reflects a structural tension in standardization: efforts to promote stability, uniformity, credibility, and reliability often conflict with the need for flexibility, responsiveness, and continuous updating. However, while greater adaptability can enhance responsiveness, it may also undermine some of the benefits that standards are designed to provide, including rigor, predictability, and stakeholder confidence derived from stable, consensus-driven processes. Addressing this tension will thus require institutional designs that support adaptation while minimizing these potential losses.

3.5 Stakeholder (Dis)Agreement and International Competition

The final cluster of challenges stems from the reality that AI governance is not merely a matter of neutral technical design. It is, instead, a *highly intensive geopolitical battleground and a stage for domestic stakeholder disputes* affecting essentially every social and economic sector.

Within individual countries, diverse stakeholder groups, including technology companies, civil society organizations, government agencies, and academic communities, often hold incompatible views on fundamental governance questions. Such debates surround topics like data, privacy, fairness, intellectual property, military use, labor impacts, transparency, the appropriate balance between innovation and precaution, and much more. An industry-led body might want to ensure that guidelines remain minimal and flexible; consumer advocacy groups might demand stringent, explicit safeguards to protect vulnerable communities. Reaching a consensus that satisfies all parties can be fraught or even impossible, particularly when stakeholders operate with unequal access to decision-making forums (Cantero Gamito & Marsden, 2024). Given the high stakes of AI systems, their dual use nature, and their impact across social sectors, the room for framing contestation and disagreement is immense (D. S. Schiff, 2022).

At the international level, the stakes are even higher, amplified by geopolitical tensions and tightly-protected national interests (Castro & McLaughlin, 2021). Countries like the United States, China, and members of the European Union each aim to shape global AI norms—whether to protect domestic markets, align with cultural values, or assert leadership in emerging technologies, both economically and militarily. Competition among these economic and political blocs can manifest as parallel standardization efforts that are at best loosely compatible, and at worst blatantly incompatible and strategically contradictory. For instance, the EU Digital Services Act imposes duties on large platforms to remove harmful content and address misinformation, while US Section 230 provides significant liability

protections to Internet platforms regarding third-party content (Citron & Franks, 2020). This regulatory divergence reflects the desire of the EU for digital sovereignty (Flonk et al., 2024) contrasted against US efforts to protect the interests of its dominant technology firms. Similar divergences arise in domains like facial recognition, social profiling and scoring, deepfake regulation, and debates over mandatory versus voluntary regulation of AI generally. These disputes emerge both across and within countries. As such, national standards bodies must navigate domestic regulatory variation and competing stakeholder interests while also taking account of international implications, and international SDOs must delicately and strategically navigate this terrain to an even greater degree. These tensions complicate ambitions for harmonization, much less anything like a single universal standard for AI ethics or governance. *Standards are political*.

Beyond these strategic conflicts, deeper structural barriers to global harmonization remain, even assuming theoretical interoperability of standards and good intentions. Organizations that operate transnationally invariably face the puzzle of reconciling local rules, national standards, and multinational frameworks that differ in scope, specificity, and strictness. Without strong multilateral bodies or treaties to unify the patchwork, genuine harmonization remains elusive (Timmermans & Epstein, 2010). Indeed, much of the harmonization to date has been limited to high-level agreement on notional principles and goals like shared prosperity, fairness, and innovation. Even the notion of "human-centered AI" might diverge across cultural or regional contexts, creating yet another fracture line that undermines fully global coherence.

Table 2. Overview of challenges for human-centered governance in a fragmented ecosystem

Challenge	Description	Key Implications
1. Translating Sociotechnical Issues into Technical Standards	Ethical constructs (e.g., fairness, well-being, transparency) are complex, political, and context-dependent, and are not easily reducible to universal metrics or technical specifications.	– Creates tension between qualitative, process-based approaches and demands for quantitative, performance-based benchmarks – Encourages oversimplification, proceduralism, and audit gaming that undermine meaningful engagement with sociotechnical risks – Risks neglecting political contestation, value pluralism, and the limits of technical codification for contested ethical concepts
2. Voluntary Nature of Standards	Most AI standards remain nonbinding, relying on consensus-based development, soft incentives, and reputational mechanisms rather than legal enforcement.	– Weak enforcement and limited accountability enable "ethics washing" and superficial forms of compliance – Gaps in coordination across internal teams and a lack of operational guidance can stall or dilute implementation – Underdeveloped feedback loops constrain collective learning, regulatory updating, and iterative improvement over time
3. Overlapping and Competing Standards	Multiple standards bodies, industry consortia, and jurisdictions have developed overlapping, duplicative, or partially	– Creates organizational confusion and decision paralysis around which standards to adopt or prioritize – Leads to redundant, inefficient, or conflicting implementation efforts across teams or geographies – Generates recursive referencing chains that form

	incompatible frameworks with little coordination.	unsustainable "compliance rabbit holes" requiring excessive navigation
4. The Pacing Problem and Difficulty in Updating	Rapid innovation in AI systems outpaces the deliberative processes typical of consensus-based standards development and regulatory reform.	– Results in standards becoming obsolete before or shortly after release, reducing their relevance and uptake – Pushes governance into a reactive posture that limits foresight, planning, and upstream risk anticipation – Undermines organizational and institutional incentives to invest in adoption, training, and capacity-building
5. Stakeholder Disagreement and International Competition	Divergent interests across governments, industry, civil society, and geopolitical blocs challenge efforts to harmonize governance frameworks and technical standards.	– Conflicting national and regional priorities contribute to fragmentation, duplication, and strategic competition – Cultural and institutional variation in the interpretation of "human-centered AI" complicates global alignment on values and goals – Concrete regulatory divergences, such as in deepfake governance, facial recognition, and content moderation, undermine the potential for interoperability and mutual recognition

A final important note is that each of these challenges—dubious technical translation of socio-technical issues, the limitations of voluntary compliance, overlapping standards, the pacing problem, and stakeholder fragmentation—interacts with the others, compounding the difficulties of creating effective AI governance. The proliferation of competing guidelines, for instance, amplifies the potential for ethics washing in voluntary regimes. Similarly, international competition heightens the pace of AI development and complicates attempts at agile updates to standards. Addressing any one of these challenges in isolation is unlikely to succeed. A genuinely human-centered governance approach must instead grapple with their interdependence, recognizing how these human and socio-technical dynamics interact with and reinforce one another. The next section therefore shifts from diagnosis to prescription, outlining institutional and design strategies that address not only individual challenges but also their overlapping effects in practice. Ultimately, *achieving meaningful harmonization will require governance structures that are themselves human-centered*: practically implementable, adaptable to organizational realities, and capable of sustaining trust and legitimacy across diverse institutional contexts.

4. Reorienting Toward Human-Centered Governance: Strategies and Recommendations

Despite the formidable challenges described above, there are pathways toward a more coordinated and genuinely human-centered system of AI standards and regulations. This section outlines several strategies designed to manage the fragmented governance landscape. *The common thread uniting these strategies is an emphasis on people—those who develop, implement, and are affected by AI systems*. Rather than viewing governance strictly as a set of technical

protocols, human-centered governance treats standards, regulations, and frameworks as tools that must be navigable, adaptable, and responsive to the realities of stakeholders working in real organizations and other settings.

4.1 Working Towards Greater Harmonization in AI Standards

Helpfully, harmonization does not require that every country or organization adopt a single, monolithic standard (Torres & Ali-Vehmas, 2024). Instead, it suggests aligning fundamental principles (such efforts are underway) and creating interlinkages between frameworks in order to lower compliance costs, reduce contradictory requirements, and foster cross-border cooperation. As outlined in Table 1, achieving this kind of harmonization involves navigating a layered ecosystem of laws, standards, certification schemes, organizational frameworks, and best practice tools, each of which contributes in different ways to the evolving AI governance landscape. In practice, realizing more coherent standards could demonstrate compliance within and across jurisdictions; reduce confusion through harmonizing vocabulary, metrics, and other best practices; and foster cooperation on research, data sharing, documentation, and policy learning. Some strategies for alignment include the following:

- **Expand scope of international dialogues, fora, and joint workshops:** Existing initiatives, such as the Global Partnership on AI (GPAI) and the OECD AI Policy Observatory, already convene stakeholders from multiple sectors and regions. Extending these dialogues to focus explicitly on standards *and their implementation*—rather than just the focus of the 2010s on high-level ethical principles—can help align technical terminology, risk assessment methodologies, and audit protocols. This means dedicated meetings and sessions working on shared agendas and getting into details, not merely sharing high-level information or objectives. For example, the EU AI Act anticipates that conformance with harmonized standards such as ISO/IEC 42001 may provide presumption of conformity for AI management systems under the Act, a particularly salient example of standards-regulation harmonization. Further, international conferences like an International AI Standards Summit and AI Standards Forum Conference (ANSI, 2024; ISO, 2025) are working deliberately toward this end, sharing insights across SDOs, industry, and regulators.
- **Formalize more SDO-to-SDO collaborations:** Bodies like IEEE, ISO, IEC, and the British Standards Institution (BSI) do sometimes coordinate through ad hoc working groups. Formalizing these collaborations under joint memoranda of understanding could eliminate redundant committees and ensure that, for example, a standard on algorithmic bias references relevant sections of a parallel emerging standard on data governance. SDOs should strongly consider opening up their processes to enable sharing of content even during the draft stage, allowing for duplication of text, and directly incentivizing membership of individuals on multiple similar standards. Organizations that encourage individuals to join SDO processes should likewise encourage them to join potentially two or more parallel standards efforts to encourage convergence. One such relevant initiative (IEEE Standards Association, 2024) is an effort by IEEE to create a joint specification for the evaluation of trustworthy AI based on IEEE's CertifAIEd, VDE's VDESPEC 90012, and the Positive AI framework, and designed in alignment with the EU AI Act. This joint specification

exemplifies efforts to unify and streamline the assessment of AI systems across Europe and globally, with the goal of promoting interoperability across regions, and between regulations and certification practices.

- **Develop policy mapping tools and standards crosswalks:** Tools that map overlaps between frameworks (e.g., the "crosswalk" approach used by the NIST Cybersecurity Framework and AI RMF) can ease confusion for practitioners, and these practices are indeed commonly used. They can help map standards and compliance requirements within and across jurisdictions or sectors. The NIST AI RMF initiative, for instance, has published crosswalk mappings that align its framework with major global standards and regulatory instruments, including ISO/IEC 42001, the EU AI Act, the OECD AI Principles, Canada's Algorithmic Impact Assessment, Singapore's Model AI Governance Framework, and the UK AI Assurance Roadmap (2025). In a sectoral context, creating working groups to establish NIST AI RMF Profiles represents an emerging and promising approach to enable systematic mapping and adaptation of standards to domain-specific needs. For example, joint dialogues between representatives of the healthcare policy subsystem in a given country and AI standards committees could foster deliberate mapping and alignment in the AI healthcare space.
- **Pursue mutual recognition agreements for standards and certifications:** In the policy realm, "mutual recognition agreements" can encourage countries or regions to accept each other's standards and certification schemes, thus lessening the burden on organizations operating globally. This strategy is highly underutilized but could reduce duplication substantially. Even in the case of incomplete agreement, it would be possible to identify which portion of a certain standard or framework is met through adherence to a peer's framework, and which additional pieces would be needed.
- **Strengthen conflict resolution mechanisms for overlapping standards:** Finally, international standardization bodies do have formal conflict resolution processes for discussing overlapping standards. However, these processes remain underused. Reinvigorating them with explicit AI-related policies—potentially with an inter-supra-national body acting as arbiter—could mitigate competition that undermines adoption. Encouragement from government, industry, civil society, and academia could push standards development actors towards more harmonization.

That said, complete harmonization may be unattainable, particularly given stark geopolitical divides and local cultural and regulatory preferences. Yet, the goal is not perfect uniformity nor uniformity for its own sake (which could minimize valuable experimentation and local relevance to human needs), but rather a level of coherence that preserves human-centered values while minimizing duplication and confusion (Uuk & Tamkivi, 2025).

4.2 Adopting Satisficing Approaches in Organizational Framework Selection

The next proposal, while unusual, may be necessary. In essence, a recurring danger in AI governance is the quest for a "perfect" standard or a single best framework, such as a standard that perfectly captures and quantifies all dimensions

of fairness. Yet the real world, AI, and the fragmented governance environment rarely offer perfection, and organizations can become paralyzed by the sheer number of overlapping guidelines.

A satisficing approach, adapted from decision theory and organizational sociology, suggests one option: pursuing "good enough" standards that meet key ethical and legal criteria, while acknowledging that no single framework will capture every nuance. This does not mean making ad-hoc, uninformed decisions, but having careful and intentional criteria behind one's satisficing. Below are a few such criteria, proposed as a starting point to help governance decision-makers satisfice thoughtfully:

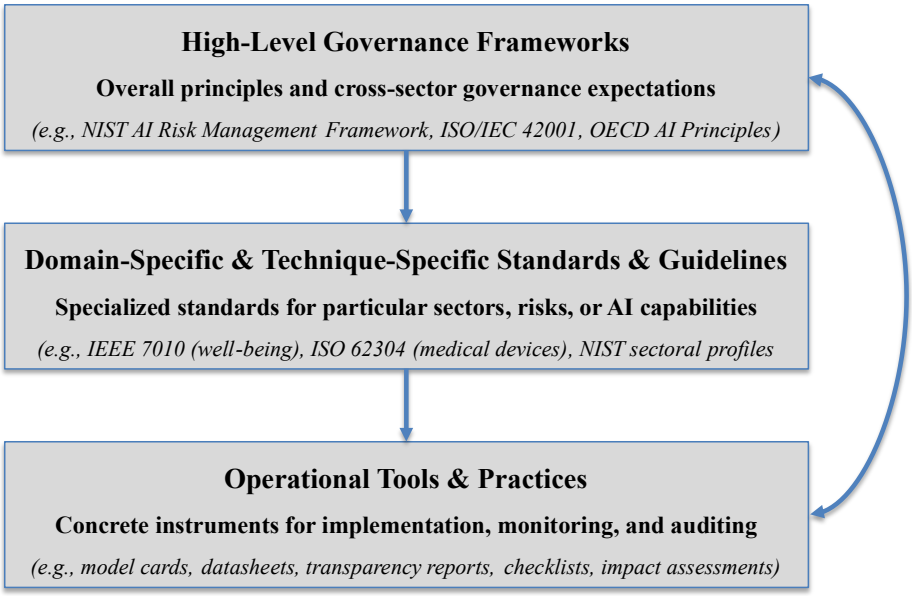
- **Identify domain-relevant standards based on sector and use case:** If an AI system is primarily designed for healthcare diagnostics, adopting specialized standards on data privacy and medical device reliability may be more urgent than implementing broad guidelines on AI transparency. This is why organizations can often first look to the relevant subject matter and regional scope when considering relevant standards, which is easier to do for specialist organizations that operate single products, for example. Guidance from professional associations, or sectoral regulatory profiles (like those encouraged by the NIST RMF) and maps can be a good starting point.
- **Engage with diverse stakeholder expectations:** Different groups—customers, citizens, civil society, government regulators—will care about different attributes of governance of an AI system (e.g., interpretability, fairness, environment impact). A bottom-up assessment of who your stakeholders are and what they prioritize can guide standard selection. However, in practice, this can lead to emphasizing the priorities of those with power. As such, developing organizations' satisficing approaches to deliberately incorporate less powerful stakeholders in their consideration is highly encouraged. Standards like ISO 26000 and IEEE 7010 explicitly build in community involvement and stakeholder engagement (Lewis et al., 2020; D. Schiff et al., 2020).
- **Prioritize enforceability and credible oversight:** Voluntary standards that lack credible enforcement can be overshadowed by even weaker but mandatory regulatory requirements. In practice, organizations tend to prioritize a standard that has regulatory support or robust third-party audit mechanisms with market-relevant signals. This has a couple of implications. One, that governments and SDOs must work extra hard to make the most prominent regulations and standards robust—well-designed, practical, flexible but resilient, and enforceable.[10] Two, that the auditing ecosystem should incorporate robust human-centered governance notions as well. That is, rather than diminishing regulations and auditing to the lowest common denominator, these key actors should incorporate highly recommended practices like meaningful stakeholder engagement, sociotechnical assessment, real-world impact assessment, and other non-minimalistic ethics and governance practices.

[10] See the argument by Schiff et al. (2021) that effective AI frameworks must be broad, operationalizable, guided and flexible, iterative, and participatory.

Next, rather than expecting one framework to address every need, organizations might adopt a *layered and sequential* adoption approach.

- **Begin with high-level guidance:** For instance, organizations might start with a primary standard, a high-level risk management or ethics framework that sets broad requirements—e.g., NIST's AI RMF or ISO/IEC JTC 1/SC 42 core guidance, or a robust sector-specific framework.
- **Add domain-specific standards and technical supplements:** Next, organizations might adopt additional domain or task/technique specific "add-ons" like IEEE 7010 (for well-being metrics) or a standard specifically for chatbots, or for medical devices, or environmental impacts.
- **Integrate operational tools and team-level practices:** In many cases, organizations will need additional detail to operationalize even these goals. As such, they should gradually adopt supplemental guidelines, team practices, software tools, and other governance tools (e.g., checklists, model cards, transparency reports). This can include gradually adopting tools that are emerging as best practice, with some consistency and robustness, as well as customizing tools or guidelines internally to translate them to specific organizational or team settings.
- Finally, in practice, note that this sequential organizational adoption may not start from the top down as depicted in Figure 3. It may alternatively start with the adoption of operational tools based on urgent needs until appropriate higher-level frameworks are identified, and such a process may be iterative.

Fig 3. Maturity Pathway for Layered Adoption of AI Governance Standards



Note: Organizations may adopt and layer these elements in varying orders depending on context, capacity, and governance maturity.

A layered approach can help organizations avoid prematurely over-engineering their compliance effort or ignoring important sectoral, organizational, or contextual nuances. It can render the adoption of human-centered AI governance less overwhelming and more feasible. To enable this, organizations should consider explicitly mapping their highest-level guidance to their lowest-level guidance, performing responsible AI maturity assessments (or similar), and laying out timelines for sequential adoption. For example, in the healthcare domain, the Coalition for Health AI (CHAI) is taking such a layered and sequential approach to responsible AI governance. Its Blueprint for Trustworthy AI in Healthcare (2023) is explicitly mapped to the NIST AI RMF and is being operationalized through working groups developing additional tools, including a responsible AI guide, a model card template, a public-facing model registry, and an inventory of assurance providers. These are all components of effective AI governance that are being developed gradually by the coalition and ostensibly adopted by member organizations.

This satisficing approach recognizes that no single framework will capture every risk, stakeholder value, or technical contingency. Rather than chasing an elusive ideal of comprehensive coverage from the beginning, a layered adoption strategy allows organizations to build reasonably robust governance structures by combining complementary standards that collectively cover key priorities. Of course, this inevitably involves trade-offs. By definition, satisficing accepts some degree of incompleteness: organizations may leave certain risks less fully addressed than would be possible under a perfectly comprehensive or fully rational system. But this reflects a realistic accommodation to humans' and organizations' bounded capacity, limited resources, and practical implementation challenges. And while using satisficing as an excuse for minimalism is a concern that warrants monitoring, such pragmatic compromise need not be especially pernicious. In many cases, organizations could, if sufficiently motivated, invest deeply in human-centered AI governance even with only a modest set of carefully chosen tools and frameworks.

4.3 Addressing the Pacing Problem

Given that AI innovations—from generative models to multimodal systems—can outpace conventional consensus-building, SDOs and regulators need more flexible mechanisms, along the lines of agile, adaptive, or flexible governance. The suggestions below aim to advance prior and evolving proposals:

- **Develop living documents and continuously updated standards:** One proposed solution is the use of living documents. That is, instead of waiting for years to finalize a standard, an SDO could release "interim" versions at regular intervals (e.g., every 6 or 12 months) as new AI techniques or risks emerge. They could even have a continuously updated online standard or toolkit, with clear versioning and dating that complying organizations can reference. This would require changes to some of the policies of SDOs. Similarly, leading benchmarks and evaluations can build this adaptability into their design using approaches like relative grading systems and dynamic digital safety labels (Ghosh et al., 2025; Judd & McGregor, n.d.) that evolve as new AI systems and evaluations, risks, or incidents become relevant and more easily measurable.
- **Establish rapid-response taskforces to monitor and update standards:** Next, actors could establish rapid taskforces, organized through SDOs or multi-lateral AI government or civil society groups. This might

involve monitoring domain specific regulations, AI developments, and proposing rapid additions or amendments to existing standards. In essence, these taskforces are the personnel responsible for creating and maintaining the living documents. For instance, Manheim et al. (2024) proposed the creation of an AI Audit Standards Board to oversee and update auditing methods and standards. Notably, these task forces need not be limited to industry or academic experts, but can include regulators, members of the public, and so on (Laux et al., 2024).

- **Expand regulatory sandboxes aligned with evolving standards:** Third, a solution which has already been advanced as the use of regulatory sandboxes, which can be aligned with existing and emerging standards (Charisi & Dignum, 2024). The premise of regulatory sandboxes is that AI developers and employers can test compliance, potentially in controlled or pilot environments, with additional legal immunity (Artificial Intelligence Act: Regulation (EU) 2024/1689, 2024). In turn, these organizations provide significant feedback regarding the efficacy of the regulations, and agree to comply with regulators and auditors ensuring heightened amounts of information. Regulatory sandboxes of this sort can be piloted in many different sectors and settings; this tool has arguably been underutilized, as it requires infrastructure, funding, and other incentives to set up successfully.

Finally, AI itself could facilitate more dynamic governance based on rapid solicitation and synthesis of new information and public feedback. Collaborative platforms powered by natural language processing could summarize feedback on proposed standards, detect changes across drafts, and flag emergent risks or gaps from new AI research or standards. There are emerging tools aimed these kinds of efforts, promoting co-production, deliberative democracy, decentralized social media platforms (Barnett et al., 2025; Birhane et al., 2022; Iversen et al., 2018), etc. However, it is important to emphasize that using AI can lead to shortcuts, like automatically summarizing feedback from the public while losing the detailed texture from their ideas and comments that can be gained and more robust participatory methods. This needs to be handled very carefully given the continued incentives for streamlining.

4.4 Enhancing Auditing and Certification Ecosystems

Many standards, regulations, and frameworks will only gain real-world traction if organizations can meaningfully demonstrate adherence to them. As noted in Section 3, such frameworks are often too abstract or theoretical for busy product teams, especially those operating under resource or time constraints. Auditing and certification processes can help operationalize governance, but only if they are supported by actionable tools, robust oversight, and credible signals of trustworthiness.

That is, SDOs, governments, and allied institutions should no longer rely on putting forward a high-level standard and waiting for others to hopefully operationalize i. Rather, these actors must begin to view implementation guidance and evidence collection as part of their core responsibility. That includes co-developing technical handbooks, sample code, organizational toolkits, and case studies that translate abstract frameworks into workflows, metrics, and team responsibilities. Incentivizing early adopters to participate in these processes and co-create best practices can help

accelerate institutional learning and public trust. And these efforts should ideally be evaluated rigorously, with deliberate plans for dissemination of lessons and best practices through trade journals, professional associations, and AI-focused formal or informal regulatory organizations. To build a more effective and trustworthy auditing and certification ecosystem, key strategies include:

- **Develop implementation guidance and practical toolkits:** SDOs, working with universities, governments, think tanks, and multisector taskforces, should proactively develop socio-technical handbooks, sample code, toolkits, and case studies that demonstrate how standards translate into concrete AI development processes. These resources should move beyond high-level descriptions, offering practical, context-sensitive examples that product teams can adopt in real-world workflows.
- **Integrate regulatory sandboxes and co-creation initiatives:** Governments and consortia can incentivize early adopters through grants, regulatory credits, or labeling schemes to participate in regulatory sandboxes. These organizations can collaboratively pilot standards implementation, generate empirical evidence, and co-develop operational guidance. Robust evaluations of these pilots should be publicly disseminated to accelerate collective learning and governance capacity.
- **Establish independent accreditation, oversight, and transparency mechanisms:** Lessons learned from other auditing regimes emphasize the importance of competent evidence gathering, independence, and clear public reporting. Governments or public-private partnerships should create accreditation bodies that certify auditors and audit organizations, minimizing conflicts of interest and limiting consulting-auditing entanglements. Auditors should disclose the specific metrics, frameworks, and methodologies used, enabling meaningful external evaluation and comparison. Organizations should publicly release appropriately redacted audit reports, including auditor credentials, audit processes, findings, and improvement plans. Professionalization of auditing practices, including ethical codes and independence standards, should be prioritized to build credibility and trust.

- **Foster continuous professional development for auditors:** Given AI's rapid evolution, auditors, AI ethicists, and governance professionals should engage in ongoing training. Continuous professional development programs, building on initial certification, can help auditors stay current with the unusually rapid technological developments, regulations, standards, and ethical norms. Creating a regime of continuous professional development, on top of one-off courses and degree programs, may be especially important for a domain like AI.
- **Promote mutual recognition of audits and certifications:** Governments and international bodies should explore mutual recognition agreements that accept robust, human-centered audit certifications across jurisdictions. If a product or organizational governance process is deemed "trustworthy" under a particular robust, human centered audit protocol, other regions, sectors, SDOs, and governments might accept those findings wholesale, or as a baseline, minimizing the need for additional auditing. Such agreements can reduce duplicative audits, lower compliance costs, and foster cross-border trust while preserving accountability.

Encouragingly, several initiatives are already helping to build the foundations of a more robust auditing ecosystem. For example, the IEEE CertifAIEd program, now being expanded through the Joint Specification initiative (IEEE Standards Association, 2024), aims to support training and certification for AI assessors and is helping to operationalize global auditing practices. Organizations such as the International Association of Algorithmic Auditors (IAAA) are building professional communities and advancing shared norms for algorithmic auditing practice, providing venues for peer learning and quality improvement. Government initiatives, such as the UK's AI Assurance Pilots and associated ecosystem development efforts (Department for Science, Innovation and Technology, 2024), are also promoting best practices and professionalization for AI auditing and assurance providers. Together, such efforts can help build the foundations of a more robust, independent, and trusted AI auditing ecosystem.

4.5 Embedding Stakeholder Participation in Standards and Governance

Another important aim is meaningfully achieving calls for diverse, interdisciplinary, and public participation in AI development and governance. Despite numerous calls for this, these goals are substantially under realized, lacking sufficient infrastructure, investment, experimentation, and learning from decades of best practices and stakeholder engagement. And as previewed above, meaningful stakeholder participation in AI is not merely a standalone or nice-to-have goal for normative or symbolic reasons, but rather a cross-cutting foundation for successful human-centered governance. It plays a critical role in grounding *framework selection* in real-world needs and values, enhancing the *legitimacy* of standards, supporting *adaptiveness* in the face of rapid change, and enriching *auditing* practices (Buhmann & Fieseler, 2023; Coeckelbergh, 2024; Deng et al., 2025). While previous sections previewed participation as an enabler of other strategies, it also warrants explicit treatment as a governance strategy in its own right. To meaningfully embed stakeholder participation, decision-makers should:

- **Broaden representation in standards development**: Occasional market testing or public feedback opportunities dominated by experts are not sufficient mechanisms. SDOs, regulators, and private sector organizations especially need to build in robust staffing and infrastructure as part of their mandatory AI development and governance workflows. As a starting point, SDOs and regulators should intentionally diversify their working groups and advisory boards by including civil society organizations, local communities, labor groups, the Global South, and historically marginalized populations. This includes lowering barriers to participation (e.g., by subsidizing travel, enabling virtual engagement, or offering stipends), and working with organizations that are successful in recruiting individuals and facilitating dialogues. While affected individuals or members of the general public could participate in traditional standards development processes, SDOs could alternatively or additionally facilitate special public engagement sessions more appropriate for these participants as compared to technocratic standards processes. Critically, these efforts should not be limited to expert representatives of the public, who do not always effectively or accurately represent their audiences (De Wit & Berner, 2009; Waheduzzaman et al., 2018), but should include community members themselves.

- **Institutionalize and empower multi-stakeholder councils and deliberative forums**: Mechanisms such as stakeholder councils, citizen assemblies, and deliberative panels (Birhane et al., 2022; Dreksler et al., 2025; Rowe & Frewer, 2000) can create structured, recurring opportunities for public input into the development and revision of standards, frameworks, benchmarks, etc. Existing efforts from the Partnership on AI and AI & Democracy Foundation offer blueprints for such institutional innovations already adapted to the complex context of AI, building on decades of participatory methods. Such mechanisms can help overcome common barriers to engagement, including low literacy, distrust of institutions, or lack of technical familiarity with AI—if carefully designed with inclusive materials and facilitation practices (Deng et al., 2023; Ray, 2023). While some countries have developed robust democratic methods, others have much to learn. Additionally, these mechanisms can be embedded formally, into government processes, or industry coalitions, or run by academics and civil society groups with the aims of reaching key decision-makers. Experimentation on both fronts is warranted.
- **Embed engagement into organizational workflows**: Participation should not be limited to consultation on high-level upstream principles. Organizations can and should build structured stakeholder input into their risk assessments, and can even build stakeholder processes into AI development pipelines (D. Schiff et al., 2020) and audit and certification processes (Deng et al., 2025). For example, standards like IEEE 7010 and ISO 26000 embed participatory elements within organizational governance expectations. That is, participation need not be limited to a single stage in the AI lifecycle, nor only to government settings. Diverse stakeholders can not merely play a role in initiating standards or advising on high-level government goals, but can also interact within organizations and design processes to ensure thoughtful implementation of those standards.
- **Invest in participatory infrastructure**: Finally, developing robust participation is neither easy nor free. Effective engagement requires sustained support: dedicated staff, training, outreach strategies, plain-language materials, translation tools, and accessible platforms (Deng et al., 2023; Kuo et al., 2024; Young et al., 2019, 2024). One-off surveys or symbolic panels are not enough—participation must be operationalized as a recurring and well-resourced function. Whether this is implemented via user experience, product, ethics, or other organizational functions is an open question that merits experimentation. And because many organizations lack incentives or know-how to do this voluntarily, mandates—through regulation or standards—may be necessary. For example, SDOs or regulators could require evidence of engagement processes as part of certification pathways or conformance declarations.

As the strategies in this chapter suggest, human-centered governance is not solely a matter of technical soundness or institutional coordination. It requires that people and communities affected by AI systems play an ongoing role in shaping the rules and values that govern them, and in facilitating human-centered implementation and evaluation of outcomes. Embedding participation meaningfully also strengthens the other approaches outlined here. It grounds satisficing strategies (Section 4.2) by revealing which standards and frameworks align with human needs and expectations. It enhances adaptiveness (Section 4.3) by surfacing emerging issues such that evolving standards remain

connected to public concerns. It reinforces auditing ecosystems (Section 44) by encouraging transparency and accountability. Yet, if calls for inclusion are to succeed, participation must be *institutionalized*, not just invited. This will require regulation, organizational mandates, and substantial resource allocation. But the payoff is considerable: standards and frameworks that are not only more politically credible, but also more resilient and widely adopted.

Table 3 summarizes the chapter's recommendations.

Table 3. Overview of suggested actions to address fragmentation and pursue human-centered governance

Strategy	Primary Goal	Recommendations
1. Harmonization of AI Standards	Reduce conflicting requirements, improve interoperability, and facilitate coordination across jurisdictions and standards bodies	– Extend international dialogues (e.g., GPAI, OECD, standards conferences) to include socio-technical deep dives on standards alignment, implementation, and audit frameworks – Expand and formalize SDO-to-SDO collaborations (e.g., IEEE–ISO joint specs) to minimize duplicative efforts and foster shared development – Create visual crosswalks, alignment matrices, and mapping tools to clarify how major frameworks (e.g., ISO 42001, NIST RMF) converge or deviate – Promote bilateral or multilateral mutual recognition of certification schemes to reduce compliance redundancy – Activate existing conflict resolution mechanisms within standards bodies to address overlapping scopes and disputes
2. Satisficing Approaches for Framework Selection	Help organizations make practical, layered, context-aware decisions amid overlapping and imperfect standards	– Use satisficing strategies to select frameworks that offer reasonable alignment with institutional priorities, regulatory expectations, and capacity constraints – Combine a general risk management framework (e.g., NIST RMF) with targeted add-ons (e.g., IEEE 7010), sector-specific guidelines (e.g., Coalition for Health AI), and practical tools (e.g., model cards, incident databases) – Select frameworks as informed by relevance to domain, geography, and affected stakeholder groups – Prioritize standards with credible enforcement mechanisms or regulatory links to avoid symbolic adoption – Use maturity models and readiness assessments to guide phased implementation across teams and projects
3. Addressing the Pacing Problem	Improve the adaptability and timeliness of governance mechanisms amid	– Maintain living documents and online toolkits with clear versioning and periodic updates (e.g., every 6 to 12 months) – Establish rapid-response taskforces to track technological change and coordinate agile revisions

	rapid technological change	– Test standards and practices in regulatory sandboxes to gather structured feedback before broader rollout – Apply NLP-based tools to synthesize input, flag emerging risks, and identify divergence across standards or jurisdictions
4. Enhancing Auditing and Certification Ecosystems	Strengthen verification infrastructure to support credible, transparent, and effective implementation of AI standards	– Provide implementation guides, templates, and sample documentation to translate abstract principles into concrete team-level practice and training materials – Establish independent oversight bodies and accreditation systems to certify auditors and prevent conflicts of interest – Publish redacted audit findings and criteria to improve transparency and learning and foster institutional trust – Encourage mutual recognition of audits across jurisdictions to reduce redundancy and support portability – Support ongoing professional development for auditors and responsible AI teams to adapt to changing standards and tools
5. Embedding Stakeholder Participation in Governance	Increase the relevance, legitimacy, and responsiveness of governance practices by embedding and resourcing inclusive and sustained stakeholder engagement	– Draw on decades of research in participatory design, technology assessment, and democratic innovation to adopt methods with demonstrated legitimacy and rigor, such as citizen juries, consensus conferences, participatory impact assessments, and advisory councils – Integrate stakeholder input throughout the AI lifecycle, not only in early-stage goalsetting or design, but also in risk assessment, deployment decisions, incident review, and post-deployment evaluation – Align participatory methods with existing organizational processes, including UX research, model documentation, red-teaming, and internal audits, to embed engagement in regular workflows – Invest in long-term participatory infrastructure, including facilitation training, digital platforms, engagement tracking systems, and dedicated staff – Partner with civic organizations, community leaders, and trained facilitators who specialize in democratic deliberation and public engagement, rather than building new systems from scratch

Conclusion

This chapter sought to problematize the proliferation of AI ethics standards, regulations, and frameworks, which has partially resulted in a tangle of overlapping, inaccessible, and impractical guidance. Despite good intentions (and some

ambitions to demonstrate leadership), issues like omitting critical sociotechnical issues when building technical standards, a reliance on voluntary mechanisms, the well-known pacing problem, and issues like geopolitical tensions are contribute into a chaotic, fragmented governance regime. These challenges collectively undermine the likelihood of achieving genuinely human-centered AI.

This chapter also seeks to contribute in a few ways. First, as a warning, it tries to robustly diagnose and characterize fragmentation driven by institutional, technological, and political forces, noting how the delegation of authority or detailing to SDOs is leading to redundancy and confusion. Second, it aims to apply the concept of human-centeredness to AI governance, not just AI. This means considering human needs and thinking about human-centered design as part of our requirements for governance institutions and ecosystems themselves. That is, how can we create standards, adopt standards, and coordinate in ways that are legible and actionable for real people in real organizations?

Fig 3. Recommendations for advancing AI standards towards harmonization and human-centered impact

To advance this effort, this chapter offers a sizable set of recommendations, some familiar and some more novel. First, it emphasizes the importance of harmonization and alignment, including the development of crosswalks, joint workshops, mutual recognition agreements, and conflict resolution mechanisms across standards bodies. Second, it highlights the practical value of satisficing approaches, encouraging organizations to pragmatically adopt and layer general, sector-specific, and tool-based frameworks based on context and readiness. Third, it addresses the pacing

problem by advocating for living documents, regulatory sandboxes, and rapid-response taskforces, alongside the infrastructure needed to support adaptive updates. Fourth, it calls for the strengthening of the auditing ecosystem through robust accreditation processes, selective public disclosure, ongoing professional training, and international audit recognition. Fifth, it stresses the need to seriously embed stakeholder participation in governance practices by drawing on decades of participatory research and institutional methods, aligning them with organizational processes, and investing in durable engagement infrastructure.

Along these lines, there are several avenues for future research (and practice) as well, displayed in Figure 3. To advance this agenda, AI governance researchers and practitioners should:

1. *Embrace empirical evaluation of standards and frameworks to assess whether they are meaningfully adopted and genuinely mitigate harms and advance human-centered goals*. While a large portion of work has gone into conceptualizing AI ethics issues and devising frameworks, only a scant portion has gone into actually evaluating the trends, challenges, and effectiveness of standards, regulations, and frameworks (Cortês et al., 2024; Fedele et al., 2024). More careful, systematic, rigorous evaluation is needed, while arguably there are enough frameworks and standards already, the overwhelming majority of which have not been evaluated in any sense.
2. *Encourage and study the development of "how-to" guides associated with standards and regulations*, such that organizations become responsible for creating actionable toolkits and navigable compliance pipelines, not just high-level standards. A baseline expectation for organizations adducing standards and frameworks is that they also be responsible for creating accompanying guides. That can involve soliciting help from other parties and gradual release plans, such as the plan by NIST to update its playbook and create sector and use-case specific AI profiles over time. But (many) organizations should no longer feels comfortable merely producing a mandate or high-level goals, lest they risk their work contributing to a problem rather than a solution. This suggests that organizations should plan their staffing, budgets, timelines, and team composition to enable guide development, real stakeholder feedback, and even longer-term goals like pilot testing and evaluation.
3. *Double down in promoting meaningful stakeholder engagement*, especially through realizing functional and scalable engagement mechanisms in government and the private sector that are built into the fabric of organizational processes. Calls for stakeholder engagement are ubiquitous, and decades of work assessing different methodologies exist to draw on. Research should move beyond calls for stakeholder engagement. While ongoing research testing different approaches in a proof of concept sense is very important, additional work is needed to focus on functional adoption of stakeholder engagement within organizational settings, and how to incentivize this..
4. *Study and refine satisficing and layering approaches,* by investigating how organizations might engage in layered or modular compliance and maturity assessments, and how this gradual adoption process could yield the greatest human-centered outcomes. Responsible AI maturity assessments are a sound starting place (CSIRO, 2022; Krijger et al., 2022), but additional educational materials and educational programs that help

organizational governance leads think through stages of adoption could be critical. Similarly, regulators, consultants, and scholars can provide structured recommendations on sequences of adoption for organizations struggling with the abundance of overlapping AI frameworks.

5. *Expand the use of sandboxes and pilots*, scaling them up to different sectors, supported by robust evaluations and well-designed iteration processes. A massive history of work on implementation, evaluation science, and analogous concepts can be drawn here and should become an increasing focus of AI governance and standards efforts.
6. *Finally, continue to promote—even insist on—sociotechnical conceptions of AI,* especially in light of the growing use of technical benchmarks. After a decade of AI governance, it appears clear that technical metrics will remain insufficient for gauging issues like well-being, human rights, and human-centeredness generally. Nevertheless, there are novel measurement strategies, evaluations, and semi-quantitative approaches that could be adduced to render sociotechnical governance more feasible (Goldstein & Sastry, 2024; Wallach et al., 2024). Substantial innovation is needed here.

In summary, some of these lines of inquiry and action may be critical to adjust course in the fragmented AI governance environment of the 21st century. Embracing human-centered design in the creation of standards and the evolution of the overall AI governance ecosystem may be necessary if we are to achieve our aspirations for AI and for our world.

References

Agarwal, A., & Agarwal, H. (2023). A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. *AI and Ethics*, *4*(2), 299–314. https://doi.org/10.1007/s43681-023-00266-9

Alan Turing Institute. (2024). AI Standards Search. *AI Standards Hub*. https://aistandardshub.org

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (arXiv:2307.03718). arXiv. http://arxiv.org/abs/2307.03718

ANSI. (2024, October 15). *Advancing AI standards collaboration: ISO, IEC, and ITU announce 2025 international AI standards summit*. https://www.ansi.org/standards-news/all-news/2024/10/10-15-24-advancing-ai-standards-collaboration-iso-iec-and-itu-announce-ai-standards-summit

Armour, J., Gordon, J., & Min, G. (2020). Taking Compliance Seriously. *Yale Journal on Regulation*, *37*, 1.

Arnold, Z., Schiff, D. S., Schiff, K. J., Love, B., Melot, J., Singh, N., Jenkins, L., Lin, A., Pilz, K., Enweareazu, O., & Girard, T. (2024). Introducing the AI Governance and Regulatory Archive (AGORA): An Analytic Infrastructure for Navigating the Emerging AI Governance Landscape. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 39–48.

Artificial Intelligence Act: Regulation (EU) 2024/1689, 2024/1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Barnett, J., Kieslich, K., Helberger, N., & Diakopoulos, N. (2025). *Envisioning Stakeholder-Action Pairs to Mitigate Negative Impacts of AI: A Participatory Approach to Inform Policy Making* (arXiv:2502.14869). arXiv. https://doi.org/10.48550/arXiv.2502.14869

Biddle, B., White, A., & Woods, S. (2010). *How Many Standards in a Laptop? (And Other Empirical Questions)* (SSRN Scholarly Paper 1619440). Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1619440&utm_source=chatgpt.com

Biddle, J. B., Nelson, J. P., & Olugbade, O. E. (2025). How Can We Know if You are Serious? Ethics Washing, Symbolic Ethics Offices, and the Responsible Design of AI Systems. *Canadian Journal of Philosophy*, 1–17. https://doi.org/10.1017/can.2025.9

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219. https://doi.org/10.1145/3351095.3372860

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. https://doi.org/10.1145/3551624.3555290

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Blösser, M., & Weihrauch, A. (2023). A consumer perspective of AI certification – the current certification landscape, consumer approval and directions for future research. *European Journal of Marketing*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/EJM-01-2023-0009

Bogina, V., Hartman, A., Kuflik, T., & Shulner-Tal, A. (2021). Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00248-0

Bogucka, E., Constantinides, M., Šćepanović, S., & Quercia, D. (2024). Co-designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 168–180.

Bolte, L., & van Wynsberghe, A. (2024). Sustainable AI and the third wave of AI ethics: A structural turn. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00522-6

Brauner, P., Glawe, F., Liehner, G. L., Vervier, L., & Ziefle, M. (2024). *AI Perceptions Across Cultures: Similarities and Differences in Expectations, Risks, Benefits, Tradeoffs, and Value in Germany and China* (arXiv:2412.13841). arXiv. https://doi.org/10.48550/arXiv.2412.13841

Brunsson, N. (2002). Standardization and Uniformity. In N. Brunsson & B. Jacobsson (Eds.), *A World of Standards* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199256952.003.0010

Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. *Business Ethics Quarterly*, *33*(1), 146–179. https://doi.org/10.1017/beq.2021.42

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*(1). https://doi.org/10.1093/ijlit/eaae011

Capel, T., & Brereton, M. (2023). What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3544548.3580959

Castro, D., & McLaughlin, M. (2021). *Who is winning the AI race: China, the EU, or the United States? —2021 update* (p. 49). Center for Data Innovation. https://www2.datainnovation.org/2021-china-eu-us-ai.pdf

Chadda, A., McGregor, S., Hostetler, J., & Brennen, A. (2024). AI Evaluation Authorities: A Case Study Mapping Model Audits to Persistent Standards. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), Article 21. https://doi.org/10.1609/aaai.v38i21.30346

Charisi, V., & Dignum, V. (2024). Operationalizing AI Regulatory Sandboxes for Children's Rights and Well-Being. *Human-Centered AI*, 231.

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, *11*(5), 545–556. https://doi.org/10.1111/1758-5899.12890

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information*, *12*(7), Article 7. https://doi.org/10.3390/info12070275

Citron, D. K., & Franks, M. A. (2020). The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform. *University of Chicago Legal Forum*, *2020*, 45.

Clouser McCann, P. J., & Shipan, C. R. (2022). How many major US laws delegate to federal agencies? (Almost) all of them. *Political Science Research and Methods*, *10*(2), 438–444. https://doi.org/10.1017/psrm.2021.32

Coalition for Health AI. (2023). *Blueprint for Trustworthy AI: Implementation guidance and assurance for healthcare*. Coalition for Health AI. https://www.coalitionforhealthai.org/

Coeckelbergh, M. (2024). Artificial intelligence, the common good, and the democratic deficit in AI governance. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00492-9

Cooper, R., & Foster, M. (1971). Sociotechnical systems. *American Psychologist*, *26*(5), 467–474. https://doi.org/10.1037/h0031539

Corrêa, N. K., Galvão, C., Santos, J. W., Pino, C. D., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. de. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10). https://doi.org/10.1016/j.patter.2023.100857

Cortês, M., Liddle, A. R., Emmanouilidis, C., Kelly, A. E., Matusow, K., Ragunathan, R., Suess, J. M., Tambouratzis, G., Zalewski, J., & Bray, D. A. (2024). AI Horizon Scanning, White Paper p3395, IEEE-SA. Part I: Areas of Attention. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2410.01808

CSIRO. (2022). Catalogue of RAI Maturity Models. *Software Systems*. https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/rai-maturity-model/

de Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? *Philosophy and Technology*, *34*(4), 1135–1193. https://doi.org/10.1007/s13347-021-00474-3

De Wit, J., & Berner, E. (2009). Progressive Patronage? Municipalities, NGOs, CBOs and the Limits to Slum Dwellers' Empowerment. *Development and Change*, *40*(5), 927–947. https://doi.org/10.1111/j.1467-7660.2009.01589.x

Delmas, M. A., & Burbano, V. C. (2011). The Drivers of Greenwashing. *CALIFORNIA MANAGEMENT REVIEW*, *54*(1), 64+. https://doi.org/10.1525/cmr.2011.54.1.64

Deng, W. H., Claire, W., Han, H. Z., Hong, J. I., Holstein, K., & Eslami, M. (2025). *WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI* (arXiv:2501.01397). arXiv. https://doi.org/10.48550/arXiv.2501.01397

Deng, W. H., Lam, M. S., Cabrera, Á. A., Metaxa, D., Eslami, M., & Holstein, K. (2023). Supporting User Engagement in Testing, Auditing, and Contesting AI. *Computer Supported Cooperative Work and Social Computing*, 556–559. https://doi.org/10.1145/3584931.3611279

Department for Science, Innovation and Technology. (2024). *Assuring a responsible future for AI* [Report]. UK Government.

https://assets.publishing.service.gov.uk/media/672a2ca440f7da695c921b7c/Assuring_a_Responsible_Future_for_AI.pdf

Dreksler, N., Law, H., Ahn, C., Schiff, D., Schiff, K. J., & Peskowitz, Z. (2025). *What Does the Public Think About AI? An Overview of the Public's Attitudes Towards AI and a Resource for Future Research* (SSRN Scholarly Paper 5108572). Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5108572

Dudley, C. (2024). The Rise of AI Governance: Unpacking ISO/IEC 42001. *Quality*, *63*(8), 27.

Emery, F. E., & Trist, E. L. (1960). Socio-technical systems. In C. W. Churchman & M. Verhulst (Eds.), *Management science: Models and techniques* (Vol. 2, pp. 83–97). Pergamon Press.

Fedele, A., Punzi, C., & Tramacere, S. (2024). The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education. *Computer Law & Security Review*, *53*, 105986. https://doi.org/10.1016/j.clsr.2024.105986

Flonk, D., Jachtenfuchs ,Markus, & and Obendiek, A. (2024). Controlling internet content in the EU: Towards digital sovereignty. *Journal of European Public Policy*, *31*(8), 2316–2342. https://doi.org/10.1080/13501763.2024.2309179

Ghosh, S., Frase, H., Williams, A., Luger, S., Röttger, P., Barez, F., McGregor, S., Fricklas, K., Kumar, M., Feuillade--Montixi, Q., Bollacker, K., Friedrich, F., Tsang, R., Vidgen, B., Parrish, A., Knotz, C., Presani, E., Bennion, J., Boston, M. F., … Vanschoren, J. (2025). *AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons* (arXiv:2503.05731). arXiv. https://doi.org/10.48550/arXiv.2503.05731

Gigerenzer, G., Reb, J., & Luan, S. (2022). Smart Heuristics for Individuals, Teams, and Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(1), 171–198. https://doi.org/10.1146/annurev-orgpsych-012420-090506

Goldstein, J. A., & Sastry, G. (2024). The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 503–518.

Goodman, E. P., & Trehu, J. (2022). AI Audit Washing and Accountability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4227350

Goodman, T. (2023). Thinking Outside the Technical Standardisation Box: The Role of Standards Under the Draft EU Artificial Intelligence Act. *LSE Law Review*, *9*(1). https://doi.org/10.61315/lselr.579

Hacker, P., & Passoth, J.-H. (2022). Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 343–373). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_17

Hernandez, J., Golpayegani, D., & Lewis, D. (2024). An Open Knowledge Graph-Based Approach for Mapping Concepts and Requirements between the EU AI Act and International Standards. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2408.11925

Howard, P. K. (2020). From Progressivism to Paralysis. *Yale Law Journal Forum*, *130*, 370.

Ibáñez, J. C., & Olmeda, M. V. (2021). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & Society*. https://doi.org/10.1007/s00146-021-01267-0

IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition* (p. 294). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html

IEEE Standards Association. (2024, November 21). IEEE Standards Association Announces Joint Specification V1.0 for the Assessment of the Trustworthiness of AI Systems. *IEEE Standards Association*. https://standards.ieee.org/news/joint-specification-trustworthy-ai-systems/

ISO. (2025, January 22). *World-first international AI standards summit to be held in 2025, announced in davos*. https://www.iso.org/news/2025/01/world-first-international-ai-standards-summit-announced-in-davos

Iversen, O. S., Smith, R. C., & Dindler, C. (2018). From Computational Thinking to Computational Empowerment: A 21st Century PD Agenda. *Proceedings of the 15th Participatory Design Conference: Full Papers - Volume 1*, 7:1-7:11. https://doi.org/10.1145/3210586.3210592

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Judd, N. C., & McGregor, S. (n.d.). *Adaptive Digital Safety Labels*.

Kijewski, S., Ronchi, E., & Vayena, E. (2024). The rise of checkbox AI ethics: A review. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00563-x

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2022). The AI ethics maturity model: A holistic approach to advancing ethical data science in organizations. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00228-7

Kuo, T.-S., Chen, Q. Z., Zhang, A. X., Hsieh, J., Zhu, H., & Holstein, K. (2024). *PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation* (arXiv:2409.15644). arXiv. http://arxiv.org/abs/2409.15644

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Three pathways for standardisation and ethical disclosure by default under the European union artificial intelligence act. *Computer Law & Security Review*, *53*, 105957. https://doi.org/10.1016/j.clsr.2024.105957

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

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787. https://doi.org/10.1145/3630106.3659005

Madaio, M. A., Chen, J., Wallach, H., & Wortman Vaughan, J. (2024). Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 214:1-214:20. https://doi.org/10.1145/3653705

Manheim, D., Martin, S., Bailey, M., Samin, M., & Greutzmacher, R. (2024). The Necessity of AI Audit Standards Boards. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2404.13060

Marchant, G. (2019). *"Soft law" governance of artificial intelligence*. UCLA: The Program on Understanding Law, Science, and Evidence (PULSE). https://escholarship.org/uc/item/0jq252ks

Marchant, G. E., Allenby, B. R., & Herkert, J. R. (2011). *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. Springer Science & Business Media.

Martínez-Plumed, F., Barredo, P., hÉigeartaigh, S. Ó., & Hernández-Orallo, J. (2021). Research community dynamics behind popular AI benchmarks. *Nature Machine Intelligence*, *3*(7), 581–589. https://doi.org/10.1038/s42256-021-00339-6

Matus, K. J. M., & Veale, M. (2021). Certification systems for machine learning: Lessons from sustainability. *Regulation & Governance*, *16*(1), 177–196. https://doi.org/10.1111/rego.12417

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Morandín-Ahuerma, F. (2023). *IEEE: A global standard as an ethical AI initiative* (pp. 127–136).

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1198–1212. https://doi.org/10.1145/3593013.3594074

Narayanan, M., Seymour, A., Frase, H., & Elmgren, K. (2023). *Repurposing the Wheel: Lessons for AI Standards*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/repurposing-the-wheel/

National Institute of Standards and Technology. (2025). *Crosswalks to the NIST artificial intelligence risk management framework (AI RMF 1.0)*. https://airc.nist.gov/airmf-resources/crosswalks/

NIST. (2023). *AI Risk Management Framework: AI RMF (1.0)* (error: NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Oesterling, A., Bhalla, U., Venkatasubramanian, S., & Lakkaraju, H. (2024). Operationalizing the Blueprint for an AI Bill of Rights: Recommendations for Practitioners, Researchers, and Policy Makers. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2407.08689

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5

Paeth, K., Atherton, D., Pittaras, N., Frase, H., & McGregor, S. (2024). *Lessons for Editors of AI Incidents from the AI Incident Database* (arXiv:2409.16425). arXiv. http://arxiv.org/abs/2409.16425

Park, S. (2024). *Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework* (arXiv:2303.11196). arXiv. https://doi.org/10.48550/arXiv.2303.11196

Paskov, P., Berglund, L., Smith, E., & Soder, L. (2024). GPAI Evaluations Standards Taskforce: Towards Effective AI Governance. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2411.13808

Pelekis, S., Karakolis, E., Lampropoulos, G., Mouzakitis, S., Markaki, O., Ntanos, C., & Askounis, D. (2024). Trustworthy Artificial Intelligence in the Energy Sector: Landscape Analysis and Evaluation Framework. *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*, 1–10. https://doi.org/10.1109/ice/itmc61926.2024.10794222

Radclyffe, C., Ribeiro, M., & Wortham, R. H. (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1020592

Raji, I. D. (2022). From Algorithmic Audits to Actual Accountability: Overcoming Practical Roadblocks on the Path to Meaningful Audit Interventions for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 5. https://doi.org/10.1145/3514094.3539566

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024). *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices* (arXiv:2411.12990). arXiv. https://doi.org/10.48550/arXiv.2411.12990

Rowe, G., & Frewer, L. J. (2000). Public participation methods: A framework for evaluation. *Science, Technology, & Human Values*, *25*(1), 3–29. https://doi.org/10.1177/016224390002500101

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2746–2753. https://doi.org/10.1109/SMC42975.2020.9283454

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81–94. IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2021.3056286

Schiff, D. S. (2022). *Setting the Agenda for AI: Actors, Issues, and Influence in United States Artificial Intelligence Policy*. https://doi.org/10.17605/OSF.IO/KW8XD

Schiff, D. S. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy. *Review of Policy Research*, *40*(5), 729–756. https://doi.org/10.1111/ropr.12535

Schiff, D. S., Kelley, S., & Camacho Ibáñez, J. (2024). The emergence of artificial intelligence ethics auditing. *Big Data & Society*, *11*(4), 20539517241299732. https://doi.org/10.1177/20539517241299732

Schiff, D. S., Laas, K., Biddle, J. B., & Borenstein, J. (2022). Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies. In K. Laas, M. Davis, & E. Hildt (Eds.), *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields* (pp. 121–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-86201-5_7

Schiff, D. S., Wilhelm, A., Schiff, K. J., & Girard, T. (2025). *The Influence of Ethics Commitments and Audits on Public Trust in AI*. https://doi.org/10.17605/OSF.IO/7CP6Z

Schirm, S. A. (2010). Leaders in need of followers: Emerging powers in global governance. *European Journal of International Relations*, *16*(2), 197–221. https://doi.org/10.1177/1354066109342922

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Seo, D., & Koek, J. W. (2012). Are Asian Countries Ready to Lead a Global ICT Standardization?: *International Journal of IT Standards and Standardization Research*, *10*(2), 29–44. https://doi.org/10.4018/jitsr.2012070103

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109–124. https://doi.org/10.17705/1thci.00131

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2025). *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence* (arXiv:2408.12622). arXiv. https://doi.org/10.48550/arXiv.2408.12622

Solanki, P., Grundy, J., & Hussain, W. (2022). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00195-z

Stahl, B. C. (2023). Embedding responsibility in intelligent systems: From AI ethics to responsible AI ecosystems. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-34622-w

Stranieri, A., & Sun, Z. (2022). A Process-Oriented Framework for Regulating Artificial Intelligence Systems: In Z. Sun & Z. Wu (Eds.), *Advances in Business Information Systems and Analytics* (pp. 96–112). IGI Global. https://doi.org/10.4018/978-1-7998-9016-4.ch005

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1), 1–20. https://doi.org/10.1038/s41746-024-01258-7

Timmermans, S., & Epstein, S. (2010). *A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization**. https://doi.org/10.1146/annurev.soc.012809.102629

Tong, R., Li, H., Liang, J., & Wen, Q. (2024). Developing and Deploying Industry Standards for Artificial Intelligence in Education (AIED): Challenges, Strategies, and Future Directions. *arXiv.Org*. https://doi.org/10.48550/ARXIV.2403.14689

Torres, A. P. G., & Ali-Vehmas, T. (2024). Governing through standards: Artificial intelligence and values. *Proceedings of the 28th EURAS Annual Standardisation Conference – Comprehensive Standardisation for Societal Challenges*. https://acris.aalto.fi/ws/portalfiles/portal/166828632/EC_Gonzalez_Torres_Ali-Vehmas_AI_standards_values.pdf

Uuk, R., & Tamkivi, S. (2025, May 20). The EU should cut actual red tape, not AI safeguards. *Fortune*. https://fortune.com/2025/05/20/eu-ai-regulations-startups/

Vallor, S., & Ganesh, B. (2023). Artificial intelligence and the imperative of responsibility: Reconceiving AI governance as social care. In *The Routledge Handbook of Philosophy of Responsibility*. Routledge.

von Ingersleben, N. (2023). Competition and cooperation in artificial intelligence standard-setting: Explaining emerging patterns. *Review of Policy Research*.

Waheduzzaman, W., As-Saber, S., & Hamid, M. B. (2018). *Elite capture of local participatory governance*. https://doi.org/10.1332/030557318X15296526896531

Wallach, H., Desai, M., Pangakis, N., Cooper, A. F., Wang, A., Barocas, S., Chouldechova, A., Atalla, C., Blodgett, S. L., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., & Jacobs, A. Z. (2024). *Evaluating Generative AI Systems is a Social Science Measurement Challenge* (arXiv:2411.10939). arXiv. http://arxiv.org/abs/2411.10939

Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109. https://doi.org/10.1613/jair.1.13196

White House. (2022). *Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age* (p. 73). White House, Office of Science and Technology Policy. https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1), 20539517231177620. https://doi.org/10.1177/20539517231177620

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2024). Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.0319e3a6

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction*, *39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Young, M., Ehsan, U., Singh, R., Tafesse, E., Gilman, M., Harrington, C., & Metcalf, J. (2024). Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday*. https://doi.org/20240428092301000

Young, M., Magassa, L., & Friedman, B. (2019). Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, *21*(2), 89–103. https://doi.org/10.1007/s10676-019-09497-z