Yale Journal of Law & Technology
Volume 28, Symposium Issue

AI Evaluation and the Standards Metaphor

Amina A. Abdu[*], **Abigail Z. Jacobs**[**]

Significant attention has been devoted to the question of how best to govern artificial intelligence (AI). In addition to legislation, many policy proposals focus on extra-legal regulatory instruments. Notably, AI evaluations provide a particularly attractive solution, imposing seemingly neutral measurements across the widespread contexts in which AI operates. Because AI evaluations are driven by a wide range of actors, their adoption as a governance tool is shifting power in AI policymaking. In particular, the companies that create AI are also key players in designing and marketing AI evaluations. This Essay examines how large technology companies and government actors conceptualize self-regulation by technology companies as a legitimate policy intervention. We note that AI evaluations are often described using the language of standards, another more established soft law regulatory instrument. Drawing on the history of standards, we discuss how AI companies leverage the metaphor of standards to describe benchmarks and evaluations in order to legitimate corporate expertise. We then examine the implications of this metaphor,

---

[*] Ph.D. candidate of Information, School of Information, University of Michigan.

[**] Assistant Professor of Information, School of Information, University of Michigan, and Assistant Professor of Complex Systems, College of Literature, Sciences, and the Arts, University of Michigan.

This work was funded in part by the University of Michigan Year of Democracy Grant. We presented an earlier version of this work at the Governing Data symposium, March 28–29, 2025. The symposium was a collaboration between the Yale Journal of Law & Technology (YJOLT) and the University of Iowa's Innovation, Business, and Law Center. We thank the participants for their feedback. We also presented an earlier version of this work at the Privacy Law Scholars Conference (PLSC), May 29-30, 2025. We are grateful to the participants for helpful suggestions. Finally, we thank the student editors at Yale Journal of Law and Technology, particularly Ryan Fore, for their feedback and assistance.

describing where it is useful in the context of AI and where it obscures important policy decisions.

Article Contents

Introduction

AI evaluation has emerged as a popular solution to govern a wide range of technologies subsumed under the umbrella of generative AI.[1]

The field of AI evaluation has been advanced by a wide range of actors: historically from academia, government, and industry, and increasingly from NGOs and nascent political movements. AI evaluations measure the performance of systems by using metrics like accuracy or reliability to measure system goals like output quality, safety risks, or system capacities. These metrics, although technical and objective-sounding, are ill-defined, represent ill-specified social phenomena, and inherit different political foundations.[2] Indeed, these evaluation efforts focus on measuring and assessing key policy-relevant risks, harms, and capabilities, such as privacy, safety, and fairness.[3]

These previously-hidden politics are betrayed by some recent events: the "Woke AI" EO; the Department of Government Efficiency's "AI-first" strategy that explicitly undercuts normal political processes; America's AI Action Plan that articulates an agenda for technical AI governance through standards to lead on national security, free speech, and "American values."[4]

---

[1] On the umbrella of "generative AI" *see* A. Feder Cooper, et al., *Report of the 1st Workshop on Generative AI and Law*, ARXIV PREPRINT ARXIV:2311.06477 (2023); on locating AI evaluations as a source of progress, *see* Laura Weidinger, Deb Raji, Hanna Wallach, et al, *Toward an Evaluation Science for Generative AI Systems*, 55 NAT'L ACADS. ENG'G BRIDGE (2025).

[2] *See, e.g.*, *id.*

[3] Shazeda Ahmed, et al., *Field-building and the epistemic culture of AI safety*, 29 FIRST MONDAY 4 (2024); Bryan H. Choi, *NIST's Software Un-Standards*, 9 GEO. L. TECH. REV. 65 (2025).

[4] Exec. Order No. 14319, *Preventing Woke AI in the Federal Government*, 90 FED. REG. 35389 (July 28, 2025), https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/ [https://perma.cc/Z4VC-VGGQ]; Makena Kelly, *Elon Musk Ally Tells Staff 'AI-First' Is the Future* of *Key Government Agency*, WIRED (Feb. 3, 2025), https://www.wired.com/story/elon-musk-lieutenant-gsa-ai-agency/ [https://perma.cc/M39V-KL3B]; Hannah Natanson, Jeff Stein, Dan Diamond & Rachel Siegel, *DOGE builds AI tool to cut 50 percent of federal*

One notable tension in AI evaluations-as-governance is that the same parties that create (and sell) AI are key players in designing and promoting AI evaluations. Evaluations are simultaneously used as marketing tools, offered as governance tools, and deployed as evidence for, or against, the need for regulatory intervention. Little attention, however, has been given to the rhetoric, practices, and values that legitimize private participation in the policy-making process.

The rhetorical force of AI evaluations as a governance tool is revealed by how they appear in policy in practice. Evaluations are a tool for development, but they are invoked in a way that implies another consensus-, deliberation-, and expertise-based process: as *standards*.

Notably, on June 3rd, 2025, the Department of Commerce announced that their primary governance arm for AI, the National Institute of Standards and Technology (NIST) AI Safety Institute would be transformed into the "pro-innovation, pro-science U.S. Center for AI Standards and Innovation," for whose work evaluations would be central.[5] Similarly, the AI Action Plan explicitly aims to "leverage the U.S. position in international diplomatic and standard-setting bodies to vigorously advocate for international AI governance approaches that promote innovation, reflect American values, and counter authoritarian influence" by becoming a leader in AI governance through evaluation.[6]

---

regulations, WASH. POST (published July 26, 2025; accessed July 31, 2025 via url); *America's AI Action Plan,* (The White House July 2025), https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf [https://perma.cc/9QFF-2CUN].

[5] U.S. Department of Commerce, *Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation*, (June 3, 2025), https://azdec.org/statement-from-u-s-secretary-of-commerce-howard-lutnick-on-transforming-the-u-s-ai-safety-institute-into-the-pro-innovation-pro-science-u-s-center-for-ai-standards-and-innovation/ [https://perma.cc/85GJ-FQX3].

[6] America's AI Action Plan, *supra* note 2, at 20. On the centrality of evaluations, *see id.* at 4 ("Ensure that Frontier AI Protects Free Speech and American Values" through evaluations) and *id.* at 22 ("Ensure that the U.S. Government is at the Forefront" for evaluations of national security risks internationally)

This Essay proposes that evaluations are often promoted in policy settings using metaphors of objective measurement, and specifically, of formal standards. Here we refer to *standards* in the sense of technical and procedural standards such as those put forward by the International Organization for Standardization (ISO), not in the sense of legal standards and rules.[7] When referring to standards, we adopt NIST's definition of documentary standards: an "agreed-upon way to carry out a technical process... developed by experts in a particular subject area, and typically approved by a recognized professional organization, which then publishes them for the world to use."[8] We emphasize three central features of this definition. First, standards are technical in nature. Second, standards are consensus-based. Third, this consensus relies on domain experts and, often, a recognized *standard-setting body* to confer legitimacy.

In this Essay, we argue that evaluations are standing in for standards as the primary soft-law tool for AI governance. We use the *standards metaphor* to describe the conflation of evaluations and standards in AI governance– and the laundering and displacement of legitimate political work that results. The metaphors used to describe AI evaluations have meaningful implications for how regulatory bodies, scholars, and beyond conceptualize the role of evaluations in policy. We argue that the standards metaphor, when applied to AI evaluations, is used to legitimize AI governance solutions and actors. As a result, authority over AI governance is relocated: from regulation to evaluations, and from domain experts—whether in consumer protection, civil rights, privacy, or beyond—to the artificial expertise of AI experts.

At stake is not whether or not 'standards' is an appropriate metaphor. Indeed, standards and evaluations in other technical fields are historically deeply intertwined.[9] But the field of AI governance is rapidly evolving and contested, as is the field of

---

[7] *See, e.g.*, https://www.iso.org/popular-standards.html [https://perma.cc/Q4XB-L7UH].

[8] Nat'l Inst. of Standards & Tech., *Documentary Standards* (2022), https://www.nist.gov/feature-stories/why-you-need-standards/documentary-standards [https://perma.cc/52F6-GKMQ].

[9] *See, e.g.*, YATES, JOANNE & MURPHY, CRAIG N, ENGINEERING RULES: GLOBAL STANDARD SETTING SINCE 1880 (JHU Press 2019).

AI evaluations. What is at stake are the critical governance decisions—what systems are safe to be used and deployed, patterns of investment and disinvestment in health, education, or government administration—that are being developed and enforced through a range of stakeholders, reshaping legitimacy, and away from typical political processes. Here we shed light on the rhetorical power of the standards metaphor and the political work that it enables—and obscures.

I. The AI Evaluation Ecosystem

AI progress (and AI hype) abound. In AI development, progress is measured through evaluations, which use specific metrics to measure AI systems' performance on specified tasks defined over specific datasets. The importance of evaluation for generative AI is multi-faceted. Rishi Bommasani and colleagues, for instance, characterize it as such:

> Evaluation gives context to machine learning models: it serves as a means for (1) tracking progress — how do we measure the performance of models and how do we design improved models... (2) understanding — what behaviors do models exhibit... and how do they perform on different slices of data... and (3) documentation — how do we efficiently summarize model behavior and communicate this to diverse stakeholders.[10]

Colloquially, developing any technology depends on evaluation to know if it works or is better than some past iteration. Is it faster at some task? Is it more accurate at some task? Did it get more clicks or make more money or get more attention on some task? Here, the pedant would rightfully note that the "at some task" is relevant but often implicit in competitions of progress— of cheaper, faster, and better.[11]

---

[10] Rishi Bommasani, et al., *On the Opportunities and Risks of Foundation Models*, ARXIV PREPRINT ARXIV:2108.07258 (2021).

[11] Melanie Mitchell, *Debates on the nature of artificial general intelligence*, 383 SCIENCE (2024) (We will discuss what is evaluated more, but note that goals for AI requires a series of assumptions about how to represent these goals. Mitchell writes:

The same is true for AI evaluation: shared tasks serve as benchmarks to compare different models, often quantitatively and with results presented in the form of ranked leaderboards that similarly abstract away the context of what the numbers represent.[12] Within evaluation, Deb Raji and colleagues describe that benchmarks stand in for "anointed common problems that are frequently framed as foundational milestones... State-of-the-art performance on these benchmarks is widely understood as indicative of progress."[13] AI evaluations provide a type of quantitative evidence made to stand in for all sorts of capabilities, risks, and behaviors.[14] The practice of AI evaluation is not necessarily to examine each system individually, but to adopt a systematic, generic approach that can be applied at scale to different systems. In doing so, evaluations make the performance of different systems directly comparable along some specific dimension.[15]

> The notion of 'intelligence' in AI—cognitive or otherwise—is often framed in terms of an individual agent optimizing for a reward or goal...this is how current-day AI works—the computer program AlphaGo, for example, is trained to optimize a particular reward function ('win the game'), and GPT-4 is trained to optimize another kind of reward function ('predict the next word in a phrase').

Of course, "intelligence" as a goal is deeply fraught and well-discussed elsewhere, and we will return to this later; AI evaluations often aim to evaluate "intelligence," which itself is often interpreted using characteristics like "how frequently was the correct word predicted.").

[12] Will Orr & Edward B Kang, *AI as a sport: On the competitive epistemologies of benchmarking, in* 2024 Proc. 2024 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 1875–1884 (2024); Paul R. Cohen & Adele E. Howe, *How Evaluation Guides AI Research*, 9 AI MAG. 4 (1988); A.M. Turing, *Computing Machinery and Intelligence*, 49 MIND 433–460, at 441, 451 (1950) (characterizing performance on a task—now colloquially known as the Turing Test—over time as a sign of progress).

[13] I. Deb Raji, Emily M. Bender, Amandalynne Paullada, Remi Denton & Alex Hanna, *AI and the Everything in the Whole Wide World Benchmark*, in 2021 35TH CONF. ON NEURAL INFO. PROCESSING SYS. 1 (2021).

[14] Melanie Mitchell, *The metaphors of artificial intelligence*, 386 SCIENCE eadt6140 (2024).

[15] Cooper et al., *supra* note 1, at 15 (explaining how the implementation and interpretation of AI evaluations is not a holistic, context-sensitive activity but is instead both generic, i.e., not system dependent, and specific, i.e., limited to whatever aspect of the system the evaluation happens to measure:

These specific dimensions range in what they can convey. Evaluations of capabilities often imply progress (like increased capacity for mathematical reasoning, operationalized as answering a set of math test questions correctly); harms that imply risk mitigation (like measuring the frequency of stereotyping language or sycophancy); or behaviors that imply the presence or absence of socially or legally undesirable outputs (like regurgitating training data or high compute need).[16] However, leading scholars also agree that current benchmarks "are not well suited to improving our understanding of the *real-world* performance and safety of deployed generative AI systems."[17]

AI evaluations offer imagined progress. Recall the headlines around one such evaluation, when GPT-4 received a passing score on the bar exam.[18] Readers will quickly identify that lawyering is different from the ability to pass the bar exam. Similarly, the job of medical doctors is different from the medical licensing exam,[19] and so on and so forth. But this type

> It is well-known that the force of legal rules depends on how they are implemented and interpreted. Many decisions are made on a case-by-case basis, taking into account specific facts and context. In contrast to this approach, machine-learning practitioners evaluate systems at scale. It is common practice to define metrics that can be applied directly to every situation (or at least a large majority of them). These metrics necessarily use a pre-specified sets of features that may leave out considerations that may be important to forming a decision that appropriately accounts for broader context.).

[16] Amy Winecoff & Miranda Bogen, *Trustworthy AI Needs Trustworthy Measurements*, CTR. FOR DEMOCRACY & TECH. (2024), https://cdt.org/insights/trustworthy-ai-needstrustworthy-measurements/ [https://perma.cc/F3TW-M3UJ] (for examples of what goes into these measurements); Hanna Wallach, et al., *Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge*, ICML (2025) (for specific generative AI evaluation examples such as those given here).

[17] Weidinger, *supra* note 1 (and citations therein).

[18] Mark Sullivan, *Did OpenAI's GPT-4 really pass the bar exam?*, FAST CO. (Apr. 2, 2024), https://www.fastcompany.com/91073277/did-openais-gpt-4-really-pass-the- bar-exam [https://perma.cc/5URG-378V].

[19] Ahmed Alaa, et al., *Medical Large Language Model Benchmarks Should Prioritize Construct Validity*, 2025 ICML (2025); I. Deb Raji, Roxana

of slippery language is a marketing feature, not a bug, of AI evaluations.[20]

The slipperiness of AI evaluations enables significant reframing: of social problems as technical AI problems; of AI experts as substantive experts in the fields they claim to evaluate; and of AI systems as functional tools that actually do what the evaluations claim to measure. Evaluation (and standards) traditionally imply the presence of substantive expertise. Yet, in the context of AI evaluations, the substance of this expertise is ill-defined, spanning beyond a single technical domain to encompass a wide range of policy issues. OpenAI, for instance, says their work on safety requires that they "anticipate, evaluate, and prevent risk" in areas like child safety, privacy, bias, and election misinformation.[21] Hugging Face's Safety Leaderboard lets you check boxes to test for "Ethics," "Fairness," "Privacy" and "Non-toxicity."[22] At Anthropic, their stated concerns are for the "continued existence of humankind," for which they have deployed a technical research team.[23] These assertions act first and

---

Danesjou & Emily Alsentzer, *It's Time to Bench the Medical Exam Benchmark*, 2 NEJM AI AIe2401235 (2025).

[20] Amelia Hardy, et al., *More than Marketing? On the Information Value of AI Benchmarks for Practitioners*, PROC. 30TH INT'L CONF. ON INTELLIGENT USER INTERFACES (2025).

[21] OpenAI, *Safety* (May 8, 2025), https://openai.com/safety/ [https://perma.cc/YV39-7WXG] (Evaluation is central to their claims about safety: "Before we share our AI with everyone, we evaluate its safety"; "Safety evaluations: We run human and automatic evaluations to ensure the model complies with our safety policies.").

[22] Hugging Face, *An Introduction to AI Secure LLM Safety Leaderboard* (Jan. 26, 2024), https://huggingface.co/blog/leaderboard-decodingtrust [https://perma.cc/ZY8F-TST7] ("The LLM Safety Leaderboard aims to provide a unified evaluation for LLM safety and help researchers and practitioners better understand the capabilities, limitations, and potential risks of LLMs.").

[23] Anthropic, *Responsible Scaling Policy* v.2.1 (Mar. 31, 2025), https://www.anthropic.com/rsp-updates [https://perma.cc/22W6-VQML](for their policies on "catastrophic risks"); Anthropic, Core Views on AI Safety: When, Why, What, and How (Mar. 8, 2023), https://www.anthropic.com/ news/core-views-on-ai-safety [https://perma.cc/2AQB-Z7CX] (Note again the emphasis on evaluation:

> Our approach centers on building tools and measurements to evaluate and understand the capabilities, limitations, and potential for the societal impact of our AI systems...

foremost as marketing rather than as meaningful assessments of model risk and performance.

Evaluations are developed in academia, industry, regulatory agencies, and beyond, and thus reflect different competing incentives.[24] Yet their quantified nature allows AI evaluations to stand in as seemingly objective measures of performance, eliding the actors and incentives behind their creation.[25] As such, they offer an opportunity for companies to advertise their technologies; for countries to assert their dominance; and for political maneuvering in favor of or against regulation.[26] That is, AI evaluation—what it is, what it stands

> We are very concerned about how the rapid deployment of increasingly powerful AI systems will impact society in the short, medium, and long term... we aim to provide policymakers and researchers with the insights and tools they need to help mitigate these potentially significant societal harms

); Anthropic, *Societal Impacts* (May 8, 2025), https://www.anthropic.com/research# societal-impacts [https://perma.cc/27FJ-SZMA] (Anthropic is very open about their intent to influence policymakers with technical research. "From examining election integrity risks to studying how AI systems might augment (rather than replace) humans, the Societal Impacts team uses tools from a variety of fields to enable positive relationships between AI and people. ... Though the Societal Impacts team is technical, they often pick research questions that have policy relevance.").

[24] Maria Eriksson, et al., *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*, PROC, AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 850 (2025); Ahmed et al., *supra* note 3.

[25] Orr & Kang, *supra* note 12.

[26] *See, e.g.*, Nestor Maslej, et al., Stanford Inst. for Hum.-Centered AI (HAI), *The AI Index 2024 Annual Report* (2024), https://aiindex.stanford.edu/ [https://perma.cc/6SHE-JFSD] (giving an overview of organizational and geopolitical competition and reporting results of model evaluations); Cade Metz & Meaghan Tobin, *How Chinese A.I. Start-Up DeepSeek Is Competing With Silicon Valley Giants*, THE NEW YORK TIMES (Jan. 23, 2025), https://www.nytimes.com/2025/01/23/technology/deepseek-china-ai-chips.html [https://perma.cc/P29N-Q8AH] (In one sentence, this article shows off all of these dynamics—comparing companies and countries, comparing model performances, contextualizing AI evaluation within the effectiveness of policy interventions, highlighting that: DeepSeek "built a cheaper, competitive chatbot with fewer high-end computer chips than U.S. behemoths like Google and OpenAI, showing the limits of chip export control.").

for—extends its reach far beyond mere technical assessment, all under a veneer of impartiality.

Consider the ARC Prize Foundation, a recently founded nonprofit with the "mission to guide researchers, industry, and regulators" through AI evaluations.[27] The foundation is clear about the multivalent opportunities of the hidden politics in evaluations (emphasis added):

> "We believe evaluations are *strategic infrastructure*, they *generate ground truth* about model capabilities and progress, which *shapes public and private sector decisions* across research, security, procurement, and policy. … ARC Prize sees benchmarks as *both technical tools and diplomatic levers*." [28]

In reflecting on the success of their agenda's adoption in the US AI Action Plan, they also explicitly engage the standards metaphor–locating benchmarks (i.e., evaluations) as the path to assert American dominance over China in standard-setting.[29]

When AI evaluations stand in as the measurement of progress, they become a tool to assert political, economic, and technological progress. The previous examples reveal AI evaluations as a proxy battlefield for both geopolitics and marketing. Similarly, AI evaluations offer a space to advance the priorities and personal agendas of the billionaires behind these companies.[30] And rather than staying within niche technical communities, AI evaluations have escaped into a public AUC-measuring contest.[31] Reporting on supposed

---

[27] ARC Prize, https://arcprize.org/ [https://perma.cc/MMA7-WQJT] (last visited December 7, 2025).

[28] Lauren Wagner, ARC Prize Foundation Statement on the US AI Action Plan, ARC Prize Blog (July 29, 2025), https://arcprize.org/blog/arc-prize-response-to-ai-action-plan [https://perma.cc/K4AV-95J6] (last visited December 7, 2025).

[29] *Id.* ("The nation that sets AI measurement standards has significant influence over the global AI ecosystem… and China is positioning itself as a dominant force in AI standards-setting. A China-led global AI benchmarking system could disadvantage U.S. firms.").

[30] Julie E. Cohen, *Oligarchy, State, and Cryptopia*, 94 FORDHAM L REV 563 (2025).

[31] AUC is a common performance metric and refers to the "area under the [receiver operating characteristic] curve." It is broadly related to metrics like accuracy or false positive rate used to compare performance.

advances in AI evaluation receive credulous mainstream headlines.[32] Among important political issues, like clarifying one's role in running the government and degree of Nazi sympathies, figures like Elon Musk will mix in bluster about the measurement of future AI capabilities.[33] Recognizing the source of the fuss—looking to why AI evaluations have escaped a technical niche area with the illusion of expertise—helps identify where and how power is being relocated.

In sum, in a rapidly shifting technical, political, and regulatory environment, AI evaluations are a core locus of activity. Beyond a niche technical field within AI, evaluations are being rapidly developed to advance both AI progress and the illusion of progress. The diversity of actors shaping the goals of this field, and the explicit social goals buried within the seemingly technical measures, have made AI evaluations an important proxy battleground for epistemic and political maneuvering.

II. Standards and the Standards Metaphor

In this Part, we propose that AI evaluations are being deployed to do the political work traditionally done via standards. Policymakers have long relied on standards to regulate technical domains, where applied expertise and nimbleness are central to good governance. In the context of AI governance, technical evaluations are being introduced in response to the call for standards, resulting in what we call the

---

[32] *See, e.g.*, Kevin Roose, *A.I. Has a Measurement Problem*, THE NEW YORK TIMES (2024), https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html [https://perma.cc/L5WR-B9TV] (This article begins with the premise that "There's a problem with leading artificial intelligence tools like ChatGPT, Gemini and Claude: We don't really know how smart they are."); Kevin Roose, *When A.I. Passes This Test, Look Out*, THE NEW YORK TIMES (2025), https://www.nytimes.com/2025/01/23/technology/ai-test-humanitys-last-exam.html [https://perma.cc/PF6N-6WN9]; Lily Jamali & Liv McMahon, *OpenAI claims GPT-5 model boosts ChatGPT to 'PhD level'*, BBC (Aug. 7, 2025), https://www.bbc.com/news/articles/cy5prvgw0r1o [https://perma.cc/Y8XV-WN77].

[33] Jess Bidgood & Nicholas Nehamas, *Social Security and Sex Robots: Musk Veers Off Script With Joe Rogan*, THE NEW YORK TIMES, 2025, https://www.nytimes.com/2025/03/03/us/politics/elon-musk-joe-rogan-podcast.html [https://perma.cc/23HM-2PB9].

standards metaphor—the conflation between evaluation and standards. We first discuss important features of standards and subsequently outline how the language of standards has been deployed to describe the adoption of benchmarks and evaluations to govern AI, despite apparent differences between evaluations and standards.

A. Standards

Standards are a part of the invisible, ubiquitous infrastructure of daily life.[34] These documented procedures govern a wide range of technical policy issues, from medical device quality to food safety to cybersecurity. However, most standards are voluntary, meaning they are not required, but adopted willingly. To take on a meaningful regulatory function, standards rely on establishing their legitimacy to those who adopt the standards and to those who otherwise have a stake in the outcome of the standardization process.

Standards gain their legitimacy through both technocratic and democratic means. First, standards have a veneer of technical and scientific objectivity: they represent the "best way of doing something."[35] Second, standards are painstakingly constructed through consensus-driven procedures. As the International Organization for Standardization (ISO) writes: "Standards are the distilled wisdom of people with expertise in their subject matter and who know the needs of the organizations they represent – people such as manufacturers, sellers, buyers, customers, trade associations, users or regulators."[36] The precise shape of consensus varies between standard-setting bodies. The ISO, for instance, emphasizes rigid voting procedures, rules for balanced stakeholder representation, and policies for addressing opposing views. Meanwhile, other standard-setting bodies like the Internet Engineering Task Force (IETF) instead advocate for "rough consensus." The composition of stakeholders also varies across settings. Around the 1980s, small consortia of private firms emerged as an alternative to traditional standards bodies.

---

[34] GEOFFREY C BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES (MIT Press 2000).

[35] Int'l Org. for Standardization, *Standards* (Accessed June 25, 2025), https://www.iso.org/standards.html [https://perma.cc/LNL4-MQQY].

[36] *Id.*

These consortia accelerated the standards-making process by bypassing multi-stakeholder deliberation and prioritizing a narrower, more homogeneous set of interests.[37] Despite these variations, consensus is a key component in the adoption of standards. Because of their voluntary nature, standards risk becoming ineffective when there are too many competing standards and no clear consensus for which one to adopt.[38] The success of standards has therefore relied on establishing legitimate procedures for achieving consensus.

B. The Standards Metaphor

Viewing evaluations as standards provides a particularly attractive promise for policymakers. Through this lens, evaluations impose seemingly neutral, expert-driven, and often quantitative measurements across the widespread contexts in which AI operates.[39] But there is an important sleight of hand here. In other domains of technology, evaluations are a key part of standards, rather than serving as standards themselves. There, evaluations are used to assess established and deliberated standards, whereas in the AI space evaluations instead precede and, rhetorically, replace agreed-upon standards. This sleight of hand is what we refer to by "standards metaphor."

We might ask: are AI evaluations like standards? Yes and no. Standardization and evaluation have long been intertwined. Designing and enforcing standards relies on strong evaluation capabilities. Many of the promises and dangers of evaluations have been addressed in the history of standards. Like evaluations, standards have been deployed as marketing tools[40] and standards-making processes have been criticized for the outsized role of industry actors.[41]

---

[37] YATES, *supra* note 9, at 241.

[38] Stefan Timmermans & Steven Epstein, *A world of standards but not a standard world: Toward a sociology of standards and standardization*, 36 ANN. REV. SOCIO. 69–89 (2010).

[39] IEEE, *IEEE Std 7001-2021*, 2021 IEEE STANDARD FOR TRANSPARENCY AUTONOMOUS SYS. 1–54.

[40] Linda Garcia, *A new role for government in standard setting?*, 1 STANDARDVIEW 2–10 (1993).

[41] Marc A. Olshan, *Standards-making organizations and the rationalization of American life*, 34 SOCIOLOGICAL QUARTERLY 319–335 (1993).

However, the metaphor of standards does not capture the current evaluation landscape, in which new benchmarks are constantly emerging and there is little consensus about which evaluations are appropriate for which kinds of assessments. This flexibility means that evaluations can be chosen strategically, effectively defanging them. Little consensus on what constitutes a good evaluation leaves significant room for gamification.

Whether or not the term "standards" aligns with evaluation practices on the ground overlooks the work the metaphor is doing. The standards metaphor clarifies the tension between how industry leaders talk about AI evaluation and how evaluation is taken up in policy. Notably, AI experts emphasize the need for evaluation. Their rhetoric admits that AI must be regulated, but it also underscores the purported legitimacy of evaluations-as governance. Furthermore, their emphasis on evaluation as a technical consideration, rather than a social or political one, implies that governance must be achieved primarily through *technical expertise*, supporting the legitimacy of self-regulation. That is, the primary technical mode through which companies are advancing self-regulation is AI evaluation, taking explicit advantage of the standards metaphor.

Through the standards metaphor, AI companies paint their self-created evaluations in the same light as rigorously specified and democratically agreed upon standards. And if evaluations are as legitimate as standards, the evaluations creators must be as qualified as the standards setters. Thus, the act of merely creating evaluations turns into "proof" that the companies are experts in AI policy and a wide range of social issues including, broadly, safety. For example, OpenAI describes itself as "leading the way in safety," offering as examples their approaches to child safety ("creating industry-wide standards to protect children"), bias ("rigorously evaluating content to avoid reinforcing bias or stereotypes"), elections ("partnering with governments to combat disinformation globally"), and beyond.[42] This approach centers technical standards—in particular, AI evaluations led by OpenAI—as the key metric for a wide range of safety and

[42] OpenAI, *supra* note 21.

policy concerns.[43] Even though safety standards across child safety, bias, and elections inherit deep political, social, and moral questions, when AI evaluation stands in for other standard-setting processes, these complex issues are reduced to mere technical problems.

III. The Standards Metaphor at Work

Metaphors have been commonly used across both the law and AI, illuminating helpful comparisons and simplifying complex concepts, but also at times obscuring important distinctions.[44] In their foundational work on metaphor, George Lakoff and Mark Johnson argue that metaphors are not a matter of language alone but also meaningfully influence thoughts and actions.[45] Indeed, metaphors have a long history of shaping both technological practices and technology policy.[46] Similarly, metaphors abound in legal discourse.

We argue that in the evolving AI governance landscape, the standards metaphor serves the concurrent functions of 1) facilitating comparisons between more and less mature,

---

[43] OpenAI, *How we think about safety and alignment* (May 8, 2025), https://openai.com/safety/how-we-think-about-safety-alignment/ [https://perma.cc/6WNB-2KZE] (OpenAI is only one such company with similarly stated goals, but their claim to social goals is open. "The mission of OpenAI is to ensure artificial general intelligence (AGI) benefits all of humanity. Safety— the practice of enabling AI's positive impacts by mitigating the negative ones—is thus core to our mission. Each failure mode carries risks that range from already present to speculative, and from affecting one person to painful setbacks for humanity to irrecoverable loss of human thriving," the latter corresponding to speculative, fantastic theories of catastrophic risk and human extinction induced by AI. As discussed in §I, this mission is achieved with technical evaluations.).

[44] *See generally*, Jonas Ebbesson, *Law, power and language: Beware of metaphors*, 53 SCANDINAVIAN STUD. L. 31, 31–39 (2008); Cooper et al., *supra* note 1.

[45] LAKOFF, GEORGE & JOHNSON, MARK, METAPHORS WE LIVE BY (Univ. of Chi. Press 1980).

[46] *See* DONNA HARAWAY, SIMIANS, CYBORGS, AND WOMEN: THE REINVENTION OF NATURE (1991); Javier Carbonell, Antonio Sánchez-Esguevillas & Belén Carro, *The role of metaphors in the development of technologies. The case of the artificial intelligence*, 84 FUTURES 145–153 (2016); Sally Wyatt, *Danger! Metaphors at work in economics, geophysiology, and the Internet*, 29 SCI. TECH. & HUM. VALUES 242–261 (2004).

codified legal spaces to address emerging governance gaps and 2) legitimating AI evaluations as a governance strategy.

A. Addressing Emerging Governance Gaps

When the law faces emerging governance gaps, metaphors enable it to adapt to new situations by facilitating comparisons with established precedents and well-understood areas of legal reasoning.[47] As Angela Condello writes, "[a]nalogical and metaphorical processes are crucial in legal discourse since law is permanently adapting to the changes in reality and, therefore, is permanently facing the challenge of classifying new objects and concepts."[48] This adaptability is particularly appealing in the face of rapidly changing technologies, such as AI.

The standards metaphor takes the well-understood strategy of using standards for technical governance, and adapts it to the uncertain space of AI governance, thus facilitating a potential path forward. Notably, standards have historically been used to govern complex and changing technological domains.[49] Historically, standards have often been a tool of soft law. Rather than acting as legally binding mechanisms, standards are frequently designed to guide behavior, providing expectations around acceptable practices. Even if compliance with standards is voluntary, they can still influence practice through market coordination. Moreover, legislation or judicial decisions may reference standards, blurring the line between soft and hard law.[50] By implying that evaluations also have these qualities, the standards metaphor is thus an effective tool for arguing against the need for regulatory intervention.

The standards metaphor can also be used to delay or defeat regulation by comparing AI to other technical domains where standards have been deployed. For example, in his testimony before the U.S. Senate Subcommittee on Privacy, Technology,

---

[47] Finn Makela, *Metaphors and models in legal theory*, 52 LES CAHIERS DROIT 397–415 (2011).
[48] Angela Condello, *Metaphor as analogy: reproduction and production of legal concepts*, 43 J.L. & SOC'Y 8–26 (2016).
[49] YATES, *supra* note 9.
[50] Raymund Werle & Eric J Iversen, *Promoting legitimacy in technical standardization*, 2 SCI. TECH. & INNOVATION STUD. 19–39 (2006).

and the Law, Anthropic CEO Dario Amodei drew on the history of airplane and automobile regulation:

> [W]e should recognize that the science of testing and auditing for AI systems is in its infancy, and much less developed than it is for airplanes and automobiles. In particular, it is not currently easy to entirely understand what bad behaviors an AI system is capable of, without broadly deploying it to users. Thus, it is important to fund both measurement and research on measurement, to ensure a testing and auditing regime is actually effective. [...] Funding measurement in turn makes these rigorous standards meaningful.

The standards metaphor establishes likeness between AI and the transportation industry and positions evaluation as an appropriate governance tool in both contexts. However, because of differences in the maturity of evaluations across these domains, Amodei establishes a need to delay AI regulation and instead invest in further technical research. Technical—and, in particular, industry—expertise is therefore rhetorically positioned as the center of AI governance, or lack thereof.[51]

Standards have also been a place where public-private partnership has thrived.[52] Of course, standards-making processes have also been co-opted by private interests toward harmful ends.[53] Powerful corporations may, for example, advocate for complex standards in order to impose higher

---

[51] There is an additional metaphor hiding here: AI "safety" as safety. The framing around which and whose concerns for risks to human life, threats to the planet, and so on is a political project. Amodei's quote, for example, draws parallels to bodily harm from cars and planes. AI safety can be defined to include a range of ideas, but also emerges from a coherent epistemic and political movement: See Ahmed et al, *supra* note 3, for a thorough accounting.

[52] YATES, *supra* note 9; Timmermans & Epstein, *supra* note 38.

[53] *See* MARTHA LAMPLAND & SUSAN LEIGH STAR, STANDARDS AND THEIR STORIES: HOW QUANTIFYING, CLASSIFYING, AND FORMALIZING PRACTICES SHAPE EVERYDAY LIFE (Cornell Univ. Press 2009).

barriers to entry on potential competitors[54] or seek to control the research used standards-setting through funding and committee composition.[55] Moreover, prior work has emphasized the limits of self-regulation in technology governance in particular.[56] This perhaps makes the metaphor relevant to an emerging AI industry with influential private actors and where many technical experts are located within industry rather than government. Industry leaders are quick to lean on this historical partnership, framing this need as even more urgent than comparable (nuclear, airline, etc.) safety needs.[57] Here, the urgency of intervention calls for speed that aligns with the driving ethos of the AI sector, and an all-hands-on-deck approach that justifies continued industry involvement in governance.

B. Legitimating Governance Solutions

Beyond addressing emerging governance questions, legal metaphors also serve a second purpose: legitimation. Metaphorical and analogical thinking shape how legal problems are conceptualized, thereby legitimizing certain legal interventions while delegitimizing others. Likening corporations to people, for example, legitimates legal protections for corporations based in human rights.[58] Critically, metaphors have both rhetorical and epistemic power, not only highlighting existing parallels between domains but also *creating* new similarities between them.[59]

---

[54] DANIEL PARGMAN & JACOB PALME, ASCII IMPERIALISM (Martha Lampland & Susan Leigh Star eds., Cornell Univ. Press 2009).

[55] Timmermans & Epstein, *supra* note 38.

[56] Cary Coglianese & Evan Mendelson, *Meta-Regulation and Self-Regulation*, in THE OXFORD HANDBOOK OF REGULATION 146–68 (Robert Baldwin, Martin Cave & Martin Lodge eds., Oxford Univ. Press 2010); Jodi L. Short & Michael W. Toffel, *Making self-regulation more than merely symbolic: The critical role of the legal environment*, 55 ADMIN. SCI. Q. 361–396 (2010); Sandra Wachter, *Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond*, 26 YALE J.L. & TECH. 671–718 (2024).

[57] Aaron Gregg, Cristiano Lima-Strong & Gerrit De Vynck, *AI poses 'risk of extinction' on par with nukes, tech leaders say*, THE WASHINGTON POST (2023), https://www.washingtonpost.com/business/2023/05/30/ai-poses-risk-extinction-industry-leaders-warn/ [https://perma.cc/7T4E-S5NZ].

[58] Ebbesson, *supra* note 44.

[59] Condello, *supra* note 48.

Scholars have convincingly argued that metaphors shape legal and policy outcomes. For example, Ryan Calo argues that metaphors can be used strategically to justify their reasoning, for example deploying the cultural imagination around robots to account for decisions around complex technologies.[60] Meanwhile, Daniel Solove contends that the pervasiveness of the "Big Brother metaphor" in privacy law limits the law's ability to deal with privacy harms that do not arise from surveillance, but from a loss of control over one's information.[61] In other words, while metaphors can enhance democratic legitimacy by clarifying reasoning around technical issues, they can also obscure important differences, thus legitimizing inappropriately importing policy solutions from one domain to another. This is further complicated in the AI context, as the language of standardization has been used to refer—and give credence—to a number of practices that do not rise to the level of uniform standards.[62]

In the context of AI evaluations, the standards metaphor has significant implications for legitimacy. Through its rhetorical power, the standards metaphor legitimizes AI evaluations. By likening evaluations to standards, they become seen as stable, agreed-upon best practices. Meanwhile, through its epistemic power, the standards metaphor reshapes how evaluations themselves are understood. Evaluations are offered to stand in for technical standards, which have long been seen as a legitimate governance tool for addressing technical policy issues.

Yet, unlike standard-setting, which has been legitimated through procedures designed to curb non-arbitrariness and establishing democratic engagement, evaluations bypass these procedures. Non-arbitrariness is often used to establish legitimacy where policy is made by unelected officials. Most notably, arbitrary and capricious review assesses whether U.S. agency decisions constitute an appropriate use of authority. Non-arbitrariness has also often been central to the legitimation of standards, whether these standards are imposed

---

[60] Ryan Calo, *Robots as legal metaphors*, 30 HARV. J.L. & TECH. 209 (2016).
[61] Daniel J. Solove, *Privacy and power: Computer databases and metaphors for information privacy*, 53 STAN. L. REV. 1393 (2000).
[62] Choi, *supra* note 3.

through regulation or adopted voluntarily. Standards organizations have achieved this by encouraging participation in the standards development process and attempting to create standards that are acceptable to all affected groups.[63] However, AI evaluations have not been approached in the same way. The standards metaphor instead works to establish evaluations as non-arbitrary through three key appeals to standards. First, by appealing to the neutral, stable nature of standards, evaluations become seen as similarly objective. Second, by appealing to the expert-driven and technical nature of standards, evaluations are taken as credible stand-ins for the values they purport to measure and govern, from safety to trustworthiness. And third, by appealing to the universality of standards, evaluations are seen as performing systematically well for all groups. (Alas.[64])

Likewise, AI evaluations lean on the democratic consensus-building procedures of standards to gesture at legitimacy without themselves embracing any of these procedures. Unlike standards, attempts toward democratic legitimacy have been minimal when it comes to the design and adoption of AI evaluations. Evaluations are not built through deliberation or careful stakeholder representation, but through ad hoc releases, often created in isolation and with little concern over conflicts of interest. In their current form, evaluations bestow legitimacy (legitimacy in practice, not normatively) without meeting governance goals. Via the standards metaphor, AI evaluations bypass the deliberative procedures and domain expertise that make standards legitimate.

In short, the complexity of the AI evaluation ecosystem (discussed in Part I) masks the selves in self-regulation, and competing incentives undergird these supposedly technical advancements. Failure to recognize the political maneuvering at play—which is obscured by the standards metaphor— will mislead any attempts at governance. Explicitly embracing the politics in AI evaluation seems necessary, but existing

---

[63] Werle, *supra* note 50.

[64] *See, e.g.*, Pauline Kim, *AI and Inequality, in* 2022 CAMBRIDGE HANDBOOK ON A.I. & L. (2022); Ninareh Mehrabi, et al., *A survey on bias and fairness in machine learning*, 54 ACM COMPUTING SURVEYS (CSUR) 1–35 (2021); Solon Barocas & Andrew Selbst, *Big data's disparate impact*, 104 CALIF. L. REV. 671 (2016).

policymaking is ill-positioned and underpowered and will continue to be so if the rhetorical and epistemic power of the standards metaphor is overlooked.

IV. Governance by Evaluation?

As companies and policymakers alike rush towards AI governance solutions, AI evaluation will continue to take on an increasingly important role. AI evaluations offer a particularly attractive solution, with apparently objective measurements and seemingly sophisticated technical solutions. Yet this picture belies the web of incentives underlying the development of evaluations and their fundamental lack of rigor.[65]

Moreover, the design of evaluations involves hidden governance decisions about what makes an evaluation useful or meaningful; about whose harms or whose safety or which capabilities are important; and about which languages, social groups or cultural norms need to be respected. While we can and should see a movement towards standards—in the sense of regulation by expert consensus, appropriate expertise, and meaningful shared governance practices—the standards metaphor risks assigning unearned authority and legitimacy to AI evaluations, and further masking their social and political commitments.

The discursive practices around AI evaluations raise important questions about the role of evaluations in governance. What is the impact of the standards metaphor? How and when should we embrace it?

A. The Good, the Bad, and the Nonsense of Evaluations

How has the standards metaphor impacted the creation of actual standards? AI standards bodies now consider one of their primary roles to be a leader in the development of AI evaluations. In the U.S., this largely falls to the Departments of Commerce (through NIST) and the Department of State, following recent executive orders and the recent AI Action

---

[65] Weidinger, *supra* note 1; Russell Brandom, *How to build a better AI benchmark*, MIT TECH REV. (May 8, 2025), https://www.technologyreview.com/2025/05/08/1116192/how-to-build-a-better-ai-benchmark/ [https://perma.cc/6WAS-B8EE].

Plan[66]. Consider the language of the US and UK groups in charge of AI standards:

> The U.S. AI Safety's Institute's (US AISI) mission is to identify, measure, and mitigate the risks of advanced AI systems ... US AISI is tasked with developing the testing, evaluations, and guidelines that will help accelerate trustworthy AI innovation in the United States and around the world – with a keen focus on helping to prevent misuse of this technology by those who seek to undermine our public safety and national security.[67]

Similarly, the first core goal of the UK AI Security Institute (until recently known as the AI Safety Institute) is to "develop and conduct evaluations on advanced AI systems."[68] Despite recent political positioning of the US and UK, comparable AI institutes across the world share a similar set of goals.[69] Yet even these efforts that eschew more rigorous standards-making in favor of softer guidance are falling short. Bryan Choi has written about how efforts towards AI standards from NIST

---

[66] America's AI Action Plan, *supra* note 4.

[67] Nat'l Inst. of Standards & Tech., *U.S. Artificial Intelligence Safety Institute, NIST* (2025), https://www.nist.gov/aisi [https://perma.cc/6X3A-GG6W] (US AISI is underneath NIST. Emphasizing the objectivity of standards, they say that their work "draws on NIST's time-tested scientifically grounded processes to facilitate the development of trusted standards around new technologies.").

[68] UK AI Safety Inst., *AI Safety Institute approach to evaluations*, AI SAFETY INST. (2024), https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations [https://perma.cc/U85N-QDZ9].

[69] Giulia Torchio & Francesco Tasin, *The Paris Summit: Au Revoir, global AI Safety?,* (2025), https://www.epc.eu/publication/The-Paris-Summit-Au-Revoir-global-AI-Safety-61ea68/ [https://perma.cc/XXB9-4432] (Describing events in 2023 and 2024, across multiple countries and the EU: "Particularly significant was the creation of the AI Safety Institutes (AISIs). While they are not technically able to regulate, these institutes cooperate with governments and industry to inform regulatory strategies, conduct AI safety research and model evaluations, and lay the foundations for international AI governance." These international convenings and AISIs had previously been led by the US and the UK; the recent deviation was led by Elon Musk and JD Vance. Regardless of the recent political showboating, evaluations are still a primary activity of these efforts.).

have been ill-suited for the fragmented AI development ecosystem.[70] A primary reason for this mismatch is that the non-standards of AI standards are a particularly contested sort. Alicia Solow-Niederman writes that the development of AI standards is "entangled in a messy, normative metaprocess of legal and sociotechnical change, bound up in market processes."[71]

In a different domain, Rachel Sachs, Nicholson Price, and Patricia Zettler have written about how the FDA has shown its hand, so to speak, by revealing the politics in their evaluation practices.[72] Sachs and colleagues describe how the FDA moved beyond their more-entrenched evaluation practices of "safety" and "effectiveness," into whether or not decisions support or impede innovation.[73] More generally, Deirdre Mulligan and Kenneth Bamberger have written about how "the reality that 'governance-by-design'—the purposeful effort to use technology to embed values—is becoming a central mode of policymaking, and that our existing regulatory system is fundamentally ill-equipped to prevent that phenomenon from subverting public governance."[74] Evaluation has always been political and non-objective, but the norms and regulatory expectations around different evaluation practices are typically more stable.

Would making AI evaluations more like standards alleviate their problems as governance tools? No, particularly given the rhetorical and epistemic forces that would be strengthened. Failure to recognize those hidden governance decisions in AI evaluations, and the ecosystem in which they are legitimized, will undermine their efficacy to govern AI in the public interest. Critical algorithm scholars have done some tracing of the political and epistemic forces at play in the AI evaluation ecosystem. Will Orr and Edward Kang, for instance, diagnose

---

[70] Choi, *supra* note 3.
[71] Alicia Solow-Niederman, *Can AI standards have politics?*, 71 UCLA L. REV. DISCOURSE 230 (2023).
[72] Rachel Sachs, W. Nicholson Price II & Patricia Zettler, *Rethinking Innovation at FDA*, 104 B.U. L. REV. 513 (2024).
[73] *Id.*
[74] Deirde K. Mulligan & Kenneth A. Bamberger, *Saving governance-by-design*, 106 CALIF. L. REV. 697–784, at 697 (2018).

part of the problem from the competitive incentives of "AI as sport":

> Such is the resilient nature of standards: they are not only resistant to change, but also nearly impossible to imagine a world without. Our view is one that understands benchmarking as a standardized evaluation process that prioritizes actionability and competition over construct validity ... [G]enerative AI prompts a need for a substantial shift towards more holistic, inclusive, and socially responsible evaluation practices that transcend the competitive epistemologies of traditional quantitative benchmarking, to encompass the variability and social impact of these systems.[75]

Shazeda Ahmed and colleagues have traced the "field-building" of the AI safety community, a community that has organized to legitimate their politics through AI evaluations via AI safety research and prize competitions.[76] These concerns are not constrained to arenas of critical analysis. Active technologists in industry, government, and outside are also actively tracking the work that can be done within AI evaluations.[77] Thomas McCoy and colleagues, for instance, advocate for reframing evaluations away from a human-centric approach, which "runs the risk of highlighting the strengths of these models—their overlap with human abilities—without revealing their idiosyncratic weaknesses."[78] They advocate instead approaching models by trying to understand how generative AI systems are "shaped by the problem they are trained to solve." This type of work reveals important

---

[75] Orr & Kang, *supra* note 12.
[76] Ahmed et al., *supra* note 3.
[77] Brandom, *supra* note 65; Mitchell, *supra* note 14; Weidinger, *supra* note 2; Patrick Altmeyer, Andrew M Demetriou, Antony Bartlett & Cynthia Liem, *Position: stop making unscientific AGI performance claims*, 2024 PROC. INT'L CONF. ON MACH. LEARNING (2024); Borhane Blili-Hamelin, et al., *Stop treating 'AGI' as the north-star goal of AI research*, ARXIV PREPRINT ARXIV:2502.03689 (2025).
[78] Thomas R. McCoy, Shunyu Yao & Dan Friedman, et al., *Embers of autoregression show how large language models are shaped by the problem they are trained to solve,* 121 PROC. NAT'L ACAD. SCIS. (2024).

epistemic weaknesses in the hype-inducing AI evaluation space; indeed, McCoy and colleagues countered that a purported "spark" of artificial general intelligence was more like the "embers of autoregression."[79]

The AI evaluation ecosystem now defines where AI governance is happening. How should we negotiate the lingering governance gap? (Is it a gap? Perhaps it is closer to what Nicholson Price calls a "negative regulatory space."[80]) Other fields have these gaps, with technical expertise, standards, and evaluations as important pieces of the governance toolkit. Ultimately, governance by evaluation is not the problem. However, facing the political and economic battlegrounds hiding in seemingly technical AI evaluation development will be necessary for meaningful AI governance. Nested political movements, sloppy evaluation practices, and intertwined incentives entrench interests in evaluations.[81] Solow-Niederman made the prediction that, by undermining claims to neutrality, "taking the politics of standards seriously may undercut their utility as governance tools."[82] No wonder that companies are so eager to obscure the politics of AI evaluation on their quest for self-regulation.

B. Bringing in Lessons from Standards

Having laid out how the standards metaphor obscures the social and political work at the heart of evaluations, we discuss how unpacking this metaphor can create a path forward for incorporating evaluations into meaningful AI governance. Like all legal metaphors, the standards metaphor is a tool. We outline how this tool can serve the dual function of first, filling

---

[79] *Id.* ("LLMs could be applied to virtually any task... one recent paper [argued] that LLMs display 'sparks of artificial general intelligence.' However, it also hinders us from understanding LLMs holistically. What we are claiming is that this aspect of LLMs has been neglected in constructing effective evaluations of their capacities."); Sébastien Bubeck, et al., *Sparks of artificial general intelligence: Early experiments with GPT-4*, 2023 ARXIV PREPRINT ARXIV:2303.12712 (2023) (the paper that McCoy et al.'s "Embers of Autoregression" based their and analysis on); Mitchell, *supra* note 11 (for a broader discussion).

[80] W. Nicholson Price II, *An Incidental Standard for Medical AI,* 8 J.L. & INNOVATION 1 (2025).

[81] Eriksson, *supra* note 24.

[82] Solow-Niederman, *supra* note 71.

an unsettled governance gap and, second, conferring legitimacy to a particular governance strategy. However, danger arises when metaphor is used to pursue the second function without the first– legitimacy is conferred, but the governance gap remains.

Using evaluations to stand in for standards overlooks important features that have been central to the success of standards. While making evaluations more like standards is unlikely to resolve the political tensions at the heart of AI governance, properly implemented standards will and ought to be a part of a long-term AI governance strategy. And, just as with other types of standards, robust evaluation will be central to the success of this work. However, technical evaluations alone cannot stand in for the features that have made standards effective for governing technology. Namely, evaluations must embrace true non-arbitrariness through careful, reasoned, and tested operationalization of concepts. This operationalization question falls both to experts – not only in AI, but also in the types of harms evaluations are trying to govern – and to those affected by AI systems. Standards have been made legitimate through domain expertise appropriate to the governance context, and through democratic deliberation both in design and post-deployment to assess whether a standard reflects desired governance goals.[83] Regulatory bodies should follow a similar path when turning to evaluations to govern AI. However, the history of standards also points us to the messy webs of incentives that threaten to undermine their effectiveness.[84] Policymakers must equally attend to these incentives when assessing how to incorporate AI evaluations into their arsenal of tools. We propose that policymakers have an important role to play in ensuring that meaningful democratic deliberation is a part of the AI evaluation pipeline. This can be achieved both by ensuring that meaningful standards precede evaluations (and not the other way around) and by establishing procedures to represent a range of

---

[83] Werle, *supra* note 50.

[84] *See e.g.*, Timmermans & Epstein, *supra* note 36, at 72, arguing that "[s]tandards' objectivity, universality, and optimality are hard won victories that can be heavily contested by third parties lobbing accusations of bias and politicization."

interests, most notably the public interest, in the consensus-making process around both evaluations and standards.

C. Beyond the Standards Metaphor

In this Essay, we have argued that the standards metaphor is primarily being deployed to confer undue legitimacy to AI evaluations and the AI industry. While unpacking this metaphor allows us to deploy lessons from standards-making productively, standardization is not a panacea to the problems of AI evaluation. Importantly, scholars and policymakers must attend to the political work embedded within AI evaluation, AI standards, and AI governance. The standards metaphor becomes dangerous when it obscures this political work but can be useful when used to illuminate it. Governance-by-evaluation would require assessing 1) whether an AI system is meeting its goals and 2) whether those goals serve broader public goals. However, the goals of purportedly 'general purpose' AI are not only contested, but are also poorly operationalized.[85] Applying the logic of standards to evaluations in their current form primarily functions to achieve political goals without changing the behavior of AI systems themselves.[86] By attending to the question of legitimacy, we must continue to interrogate who has the authority to evaluate and ultimately to govern AI.

---

[85] Hardy et al., *supra* note 20; Winecoff & Bogen, *supra* note 16; Eriksson, *supra* note 24.

[86] Cary Coglianese, *The limits of performance-based regulation*, 50 U. MICH. J.L. REFORM 525 (2016) (on the shortcomings of performance-based regulation in settings where performance is underdefined).