AI & SOCIETY
https://doi.org/10.1007/s00146-025-02320-y

REVIEW

The necessity of AI audit standards boards

David Manheim[1,2] · Sammy Martin[3] · Mark Bailey[4] · Mikhail Samin[5] · Ross Greutzmacher[3]

Received: 9 February 2025 / Accepted: 13 March 2025

Abstract
Auditing of AI systems is a promising way to understand and manage ethical problems and societal risks associated with contemporary AI systems, as well as some anticipated future risks. Efforts to develop standards for auditing artificial intelligence (AI) systems have therefore understandably gained momentum. However, current approaches are not just insufficient, but can be actively harmful. Transparency alone does not address concerns about risk. Internal auditing is insufficient, and easily becomes safety-washing. External audit is better, but requires credible standards. Industry-led approaches to building standards or to perform audits lack credibility and undermine other efforts. Regulation often is ill adapted and becomes a static barrier. Lastly, all of these limited technical, governance, and even ethical assessments fail to ensure continued stakeholder input and engagement. Instead, the paper proposes the establishment of an AI Audit Standards Board, in line with best practices in other fields, including safety-critical industries like aviation and nuclear energy, as well as more prosaic ones such as financial accounting and pharmaceuticals. This would address the evolving nature of AI technologies, help maintain public trust in AI, and promote a culture of safety and ethical responsibility within the AI industry. By ensuring audits remain relevant, robust, and responsive to the rapid advancements in AI, auditing AI will not devolve into safety washing and addresses risks and ethical concerns that will continue to arise as AI becomes increasingly important in society, and as human interaction with these systems changes over time.

Keywords AI governance · AI audit · Technology audits · Ethical AI · AI policy · Organizational culture · Audit standards · Standards setting · Responsible AI

1 Introduction

Audits are used in different domains both for giving an account of what is happening within a system, and verification of requirements (Courville et al. 2003). When considering how to perform audits in a new domain, drawing on best practices from other domains is critical—and artificial intelligence (AI) is a new domain with new risks (Hendrycks 2025). There is now significant focus on audits for ensuring the safety and evaluating the risks and harms of AI systems (Mökander et al. 2023; Shevlane et al. 2023; Sharkey et al. 2023), as well as significant earlier work on audit methods for evaluation of societal implications (Raji et al. 2022a) and on what a mature ethics process involves (Krijger et al. 2023). While it is encouraging to see action addressing the critical role of evaluating and auditing risks from frontier models and on evaluation of ethical standards, there are many challenges for these types of evaluations and audits (Courville et al. 2003).

✉ David Manheim
davidmanheim@technion.ac.il

Sammy Martin
sammy@transformative.org

Mark Bailey
mark.m.bailey@ni-u.edu

Mikhail Samin
ms@contact.ms

Ross Greutzmacher
rossgritz@gmail.com

[1] Technion Israel Institute of Technology, Haifa, Israel

[2] Association for Long Term Existence and Resilience (ALTER), Rehovot, Israel

[3] Transformative Futures Institute, Wichita, KS, USA

[4] National Intelligence University, Washington, D.C., USA

[5] AI Governance and Safety Institute, Berkeley, California, USA

Published online: 05 May 2025

Below, we argue that the current approach of standards development for AI systems is harmful, among other reasons, due to the proliferation of inconsistent and rapidly outdated static standards and a lack of clarity about what is appropriate in any given domain, for any specific AI system, and for specific applications. This fragments efforts, undermines efforts to make any specific auditing methods standard, and reduces the usefulness of standards development.

To supplement important technical approaches being developed for auditing AI systems, we also need broader audits, and ongoing development of standards—an audit standards body, not just recently proposed audit standards, such as suggested by Faveri et al. (2025). To explain what is needed and why, we review past work and current audit approaches. We then explain why current efforts fail at addressing relevant challenges and risks. We also note that auditing standards are not the same thing as standards for audits, and neither necessarily implies regulation. In addition, different methods, audit approaches, and standards are needed for different model types and applications (Frase 2023). This is especially true when specific standards are unclear or disputed or when detailed standards would be unwise, as we will explore. Therefore, the body of the paper revisits some known ideas in AI audit, as well as some drawbacks of some of the approaches methods, both to provide a brief overview of the issues that evaluation and auditing should ideally address, and to show how the suggestion of audit boards differs from other approaches to standards.

1.1 Background

AI auditing refers to the process of evaluating AI systems for safety, fairness, transparency, and compliance with ethical and technical benchmarks. This includes internal audits conducted by developers, external audits by independent entities, and red-teaming to identify risks. In contrast, AI standards are guidelines and best practices that define what AI systems must do, or what vendors must do with the models, and AI audit standards are standards for how AI audits should be conducted, ensuring consistency or appropriately adapted methods for auditing different models and applications. Either class of standard may be developed by industry groups, academic researchers, or regulatory bodies. Regulatory oversight, meanwhile, is the enforcement of legal and policy frameworks governing AI systems, ensuring compliance with broader societal and governmental requirements. This type of oversight can involve audits, but while audits and standards inform best practices, regulatory oversight determines legal obligations and consequences.

Auditing is well established in the context of computing generally (van Biene-Hershey 2007, Hall and Hazell 2015). Standards, such as COBIT, date at least to the 1990s, well before the current classes of artificial intelligence. Recently, debate about safety, misuse, and bias has led to internal and external checks on models, and there are now AI systems audits, including both a growing ecosystem of AI ethics and accountability audits, Birhane et al. (2024) as well as safety audit efforts such as the red-teaming performed for GPT-4 (OpenAI 2023a, OpenAI 2023c, OpenAI 2024).

Self-reporting via "Model Cards" (Mitchell et al. 2019) has become commonplace, and internal auditing and red-teaming have become more common prior to frontier model release, with notable exceptions (Anil et al. 2023; xAI 2024). There have also been broader mandates proposed for auditing AI systems (Koshiyama et al. 2021; Faveri et al. 2025), though a principles-to-practice gap exists, especially noted for AI ethics (Tidjon and Khomh 2023), which is even more of a problem for the newer auditing methods for safety.

In addition to voluntary, oft ignored self-audits, there has been discussion of mandatory external and independent auditing, which would help address the principles-to-practice gap (Courville et al. 2003). Efforts are largely country-specific, and are likely to create a regulatory overlap failure modes seen elsewhere (Robb et al. 2023), but there are at least attempts to make standards. The now-defunct Biden U.S. Executive Order (Biden 2023) required that National Institute of Standards and Technology (NIST) work on "Developing Guidelines, Standards, and Best Practices for AI Safety and Security," following their Risk Management Framework (NIST 2023), and their mandate includes "an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities..." (Biden 2023). Similarly, the UK's AI safety institute views "publicly accountable evaluations of AI systems" to be a key part of its mission (DSIT 2023), and efforts in France have been proposed as well (Commission L'intelligence Artificielle 2024).

However, these all focus on evaluation of models themselves. Despite limitations, the efforts enable audits to be useful and successful, but they do not address broader issues, including the need for a regime that will ensure robust standards for a quickly evolving ecosystem and uncertain but rapidly evolving ethical challenges, risks, and future threats. As recent work has shown, even within AI accountability auditing, rather than the nascent and less well developed field of AI safety audits, the focus is irresponsibly limited to only model evaluations, and "this ecosystem [which includes broader ecosystem and product audits,] is muddled and imprecise." (Birhane et al. 2024).

1.2 Current practice

While efforts mentioned above for AI ethics are widespread, if fragmented, frontier AI developers now believe that their models may soon be capable of causing direct harm on a wide scale (OpenAI 2023c; Anthropic 2023). OpenAI's preparedness framework identifies cybersecurity, chemical,

biological, radiological, and nuclear (CBRN) threats, and persuasion and model autonomy as categories of dangerous capabilities that they expect from near-future models. While these dangers do not obviate the need for addressing still often ignored ethical issues, to the extent that these anticipated AI systems are high-risk, they must be treated in line with the best practices for other safety-critical settings such as aviation or the nuclear industry.

Therefore, we consider existing industry best practices, adapting them where relevant to AI evaluation, and the role of a standards board. We also discuss the need for a shift towards safety culture, not just formal standards and practices, something that is well established in safety-critical industries, but severely lacking in the AI Industry (Manheim 2023a).

In aviation safety culture audits are common and United States Federal Aviation Authority's (FAA) assessments include employees' perception of safety culture and self-assessment of how important safety is to their work. Similarly, the nuclear industry highlights the importance of regulatory oversight of safety culture, rather than only individual safeguards (IAEA 2013). In other safety-critical industries like manned spaceflight (Keyser 1974) we see management structures changed to give increased veto power to lower-level decision-makers should they notice problems.

These governance norms can be adopted by AI, not just operationally but also culturally. For example, the financial sector provides a case study on how to effectively audit and disclose risks, allowing external auditors comprehensive internal access, ensuring transparency while maintaining confidentiality. Raji et al. (2022a) and Schuett (2024)'s calls for internal audit, but aligning with Raji et al. (2022b), it is clear that a standards board for creating norms and best practices for outside audit can ensure that independent AI audits can be conducted without requiring companies to publicly reveal proprietary information.

Lastly, the pharmaceutical industry's explicit commitments to ongoing audits and publishing results can guide AI auditing. The US Food and Drug Administration (FDA), for instance, mandates regular pharmaceutical inspections with publicly available results, ensuring continuous compliance. This is in stark contrast to AI audits, currently performed on ad-hoc bases without external monitoring. Pharmaceuticals also provide a model for prioritizing oversight and resource allocation based on risk (Lawrence and Woodcock 2015). It also provides another warning about the costs and unnecessary duplication and consumer harm from differing and inconsistent national standards, with outsized harms to low and middle income countries (Moscou and Kohler 2018), as well as regulatory capture of standards and regulation (Carpenter 2013).

Most fundamentally, having a body with a mandate to produce and maintain auditing standards is essential, and in the cases discussed (for nuclear energy, aviation safety, accounting, and drug safety,) some independent body exists ensuring that these standards are not merely industry practices, which are particularly liable to regulatory capture, but widely adopted public standards. Learning from these examples for AI, without falling prey to the various failure modes which are seen in those domains, is both possible, and beneficial for all parties involved.

1.3 Problems specific to AI development

There are a number of distinct problems faced by those attempting to develop audit standards for AI ethics and safety. Before explaining the three-pronged approach an Audit Standards Board would need to take, we outline four challenges. Each of these, we feel, is poorly addressed by the current assumption that standards development will be helpful in limiting either speculative risks or current harms, and would be assisted by a different structure.

First, standards are often created and adopted in advance of an AI system's development, which sharply limits their ability to ensure that AI systems are safe prior to further training or eventual deployment. Standards produced today are unlikely to be sufficient in a year, much less several years, and they aim at a moving and unpredictable target (Zhou et al. 2023). AI system development, testing, and deployment is an iterative process that can involve a wide variety of changes. For example, ChatGPT moved from GPT 3.5, to GPT-4, to GPT-4V, to include Dall-E 3 integration, completely changing first the underlying models, then the modes of interaction, all in less than a year. The same type of radical transition happened again the following year, moving from GPT4o to o1 to o3. These transitions can also change access types, from UI access to API access, to fine- tuning and customized models. A safety or impact audit standard which would apply to ChatGPT when released in mid-2022 would be completely insufficient by the end of 2023, much less today.

In addition to the changes in systems over time, developing standards is challenging due to the emergence of capabilities in foundation models,[1] and Bommasani et al. Bommasani et al. (2021) suggest that dangerous and/or difficult to anticipate or forecast capabilities can emerge with increasingly large AI systems, when systems are scaled up by orders of magnitude (Wei et al. 2022). The result is that

[1] As defined elsewhere, foundation models are large AI models trained on one or more modalities of data in an unsupervised or semi-supervised fashion that are able to generalize well to a wide variety of downstream tasks as they are well-suited for zero-, one-, or few-shot learning and for transfer learning, i.e., transferring the knowledge learned on large corpora to downstream tasks via fine-tuning (Bommasani et al. 2021).

both accountability audits focused on system outcomes, and safety audit efforts like red-teaming, are rendered obsolete. This suggests that many proposed metrics and evaluations are likely to end up insufficient or misleading when applied to increasingly large systems. (Corso et al. 2023) One example is the emergence of language modeling in diffusion models, which can generate increasingly valid text inside of generated images, which may make audit questions previously limited to LLMs, such as biased language, necessary in image generation models as well. Furthermore, while increased model size often enables greater capabilities, it is not determinative, and can change based on applications or new methods, as occurred with reasoning models, so that even when we know what possible capabilities we are concerned about, the specific capabilities to emerge from a new foundation model are not obvious in advance.

A third issue is that evaluations of capabilities may fall short if they fail to capture new modes of human interaction with the models, especially due to emerging misuse opportunities, or via extensions of models into new systems such as reasoning models or other application-specific interfaces or versions of the model. At a minimum, extensive red teaming is necessary, even if automated (Zhu et al. 2023), but the risks associated with increasingly complex and capable models will also require novel strategies of risk assessment. There are currently significant unsolved issues with model adversarial robustness, and it is also possible to use highly capable models to adversarially attack each other, making them yet more vulnerable (Shah et al. 2023). There are also new threats, such as the recently identified "Sleeper Agents," which have emerged (Hubinger et al. 2024).

A fourth reason that standard requirements can fall short is that models are deployed in rapidly changing technological contexts. A model capable of programming that poses no risk of misuse for bioterrorism may later pose risk as biological design tools become more capable (Sandbrink 2023; Mouton et al. 2023). Similarly, new applications and uses of the model itself can create danger. Foundation models that only output text were originally assumed to be incapable of multi-stage projects and agent-like behavior. The assumption was correct, until the early 2023 creation of AutoGPT (and similar programs) that harnessed multiple instances of an LLM to pursue multi-step plans. This risk was speculative, and for that reason no internal audit of the model contemplating current uses foresaw or tested for this risk in advance. And future changes, such as multi-agent cooperation, which are currently partly speculative challenges, seem poised to occur, and will require additional caution (Hammond et al. 2025).

Another risk of static standards for evaluation of models is leakage (Kaufman et al. 2012), where data about the test set, in this case, the analyses or tests which are part of the standard, are included in the training data, or are known otherwise. The recent example of FrontierMath, where OpenAI secretly had access to Epoch AI's questions and many answers, illustrates how even without use of the test for training, selection effects can invalidate a benchmark (Meemi 2025).

Therefore, to achieve the intended evaluation and mitigation of risks and harms, we need not just standards, but a group that can do more than just evaluate and monitor the models that are produced. We divide the identified needs into three areas. First, we need to audit the process, not just the product. Second, we need to change the culture of safety complacency. And finally, we need to empower these groups, not create standard auditing procedures. Between these needs, auditing standards boards emerges as an obvious solution.

Figure 1 depicts these fundamental principles and practices that should be adhered to for each of these three elements of our proposed approach.

2 Audit the process, not just the product

As explained, a language model or other AI system poses risks that change over time. Furthermore, black-box systems also cannot necessarily be evaluated only as a final product, for instance, because training data attacks can be imperceptible in the final model (Khalid et al. 2019) making evaluation of inputs and processes necessary. We see a parallel between process evaluations for AI and cases such as pharmaceuticals and nuclear power, where the entire design process must be overseen. In all of these industries, it is impossible to tell just from looking at the final result whether it is safe and effective.

There are also risks and issues that are directly about inputs or training, such as ethical audits of data sources and human curation (Miceli and Posada 2022), and limitations and issues with RLHF for which audit standards have already been discussed (Casper et al. 2023). Moreover, an AI system is not a single static artifact that can be evaluated in isolation. The increasingly interactive use of multiple models within AI systems also means that there are complex relationships between individual changes and ethical and safety concerns. Even beyond that, multiple AI systems interacting outside the framework of a single agent will pose additional challenges (Hammond et al. 2025). An audit that addresses these concerns must be broader than the evaluation of a single product, and incorporate a human-centric view (Shin and Shin 2023).

There are also specific model evaluation techniques that require integration with the process. For example, use of predictions to evaluate model capabilities, and judging the safety of a model in relation to whether it does more than expected (or less) requires a clear process for pre-training

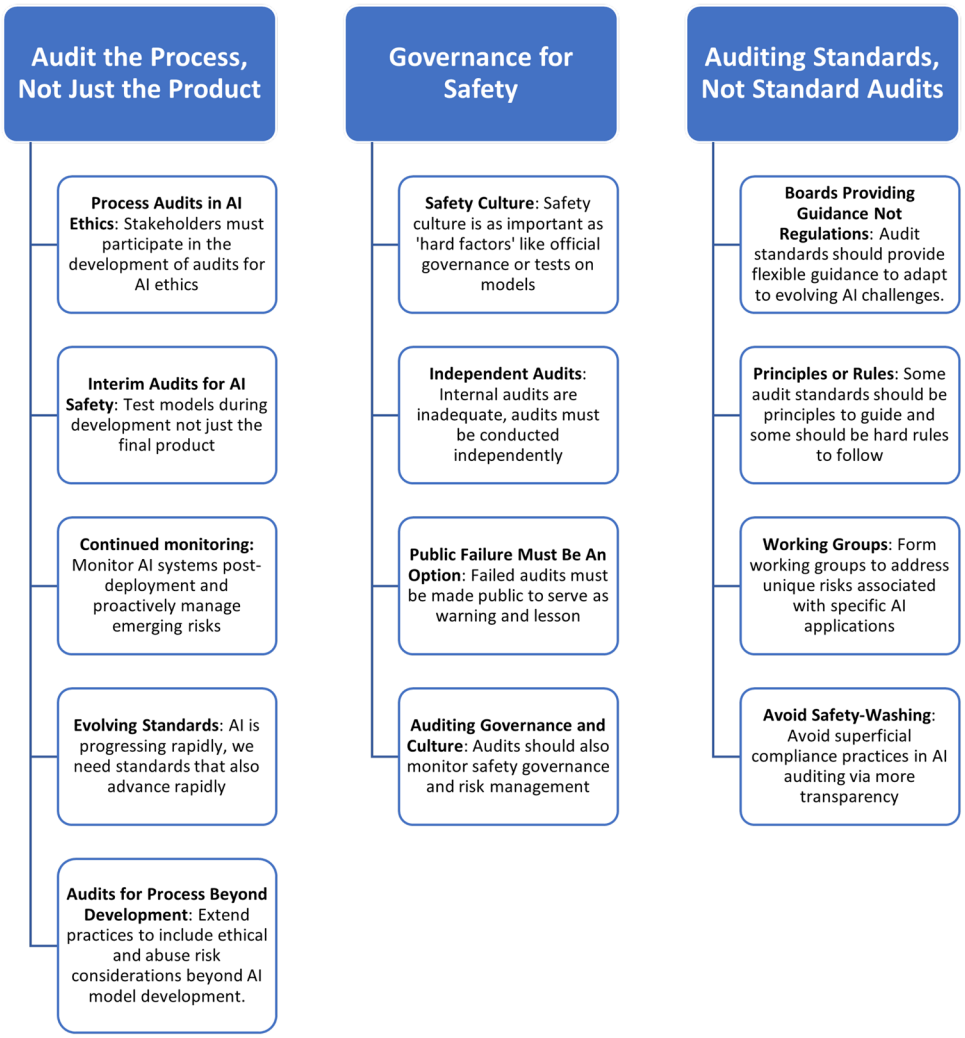

Fig. 1 We highlight the three elements of the approach we suggest that a standards board should take, and further identify best practices relevant to each that we recommend be promoted in the approach of boards, or other equivalent standards organizations

statements of predictions, and post-training evaluation. This could also address problems raised by Raji et al. (2022a), where AI projects fail to correctly do what they were designed to do.

2.1 Process audits and stakeholder participation

Model auditing involves and impacts a number of stakeholders, including model developers, auditors and safety evaluators, government and regulators, and the public. Ensuring all parties' views and needs are represented in the development of audit processes requires not just consultation with all parties, but continuing engagement (Anderljung et al. 2023b).

Raji et al. (2022a) suggest viewing audits of machine learning models as a multistage process that in part is focused on generating artifacts which can provide insight to external stakeholders. This nicely highlights both the need for process, which we discuss below, and the way that the needs of stakeholders are integral to what an audit accomplishes. While that work is focused on a subset of the concerns we discuss, the suggestions are critical, and points to exactly the sort of stakeholder engagement we argue is more broadly valuable.

That work, and the much broader literature, focuses on current harms that exist from building and using the systems as intended. These ethical issues have been highlighted by Mitchell et al. Mitchell et al. (2019). To address these harms, there is a nascent set of identified dimensions (Krijger et al. 2023) and best practices for ethics in AI. That said, this requires more than auditing the model which is created. For example, addressing harms that emerge from biased training data or appropriation of copyrighted data requires engagement with the data collection and training process, rather than just the final product. Such data audits are defined, but often remain outside of the scope of model audits (Birhane et al. 2024).

Given the current state of practice and how far short of long-standing and well understood standards for transparency almost all AI models fall (Liang 2023), these recommendations must be adopted, and failures to do so must be

highlighted. But the different equities involved mean that adopting recommendations is not enough. Instead, both those benefiting from the model and those harmed by it need to be represented when discussing what should be audited during these processes.

Furthermore, systemic and structural risks, and risks from misuse, require a sociotechnical approach to both AI safety (Lazar and Nelson 2023) and model auditing (Weidinger et al. 2023). Those harmed by algorithmic injustice are rarely represented in current efforts towards standards development, and the types of effort needed to include these stakeholders (Lombe and Sherraden 2013; Bhaumik et al. 2015) seems largely absent in AI Audit standards development, and in the research needed to enable it, respectively (Birhane et al. 2024)

In addition, as of now, it is unclear where the responsibility lies for harms caused in ways that are not directly dependent on AI systems' capabilities or accidents resulting from them; the interaction between the AI system, the deployment, and the use of the systems requires ongoing processes to understand risk. Having multi-stakeholder engagement in defining standards and requirements benefits both society as a whole by reducing risk, but it is also in the self interest of AI developers as it ensures that they have clearly defined responsibilities.

2.2 Audits for process beyond development

The inclusion of stakeholders in standards for auditing is critical, but the audits themselves must be focused on more than just the models. As Raji et al. (2020) argued, auditing needs to be present throughout the development of a model. However, auditing for harms needs to extend beyond the development of a model (Raji et al. 2022a).

Process audits can and should start with evaluating the company, its structure and governance, and its safety and ethical orientation and commitments, as discussed below, and must extend from the process of data collection and training through post-deployment monitoring. These require both ongoing checks, and revisions of what practices and norms should be monitored. For example, we can consider abuse risks, which are already occurring, that need to be addressed. Checking whether models can be abused is critical but it is probably impossible to remove all potential for abuse Therefore, ongoing monitoring and processes for auditing that work will be critical.

The abuses to address will not only be those that are occurring already. For example, mass spear-phishing is an existing threat which can leverage LLMs to further reduce costs (Hazell 2023), but waiting for proof that this type of abuse is happening before including the risk in evaluations turns risk evaluation into a reactive and ultimately ineffective process. And despite sensationalist presentations of capabilities, some types of AI-enabled bioterrorism are currently implausible (Mouton et al. 2023), but capabilities may continue to progress. For that reason, it seems likely that by the time it is clear such uses are possible, mitigations need to already be in place. Any process for audit that attempts to address any of these risks will need to move beyond point-in-time evaluation and passing or failing an audit, and require processes for ongoing risk analysis that include routinely revisiting questions about forms of abuse and the safeguards needed.

Societal impacts are hard to determine a priori, and will change as these systems develop new capabilities, as society adapts to them, and as the digital landscape changes over time. The set of concerns that should be highlighted in audits again argues for a diverse set of stakeholders, but also requires broader and continuing engagement and public discussion of how and what audits must cover, rather than one-time consultations typical of regulation or standards development. This shows that, while sociotechnical approaches and involving broader stakeholder groups are critical (Lazar and Nelson 2023), those approaches need far more than static evaluation to be meaningfully effective in reducing harm. As we argue, inclusive boards are a solution that incorporates the advantages and which can mitigate many of the problems.

2.3 Interim audits during AI training

In addition to the ethical harms and risks of abuse of models for current and anticipated risks, there are also risks arising from capabilities that AIs may possess in the near future. Several groups have noted that auditing model checkpoints and audits during training may be necessary, both because the AI's capabilities may advance suddenly during training, resulting in the need for new safeguards to be developed or applied, and because auditing the non-final version (Avin 2023; Anderljung et al. 2023a; Sharkey et al. 2023) can reveal or help predict latent capabilities.For example, safety measures like RLHF, which are essential, can nonetheless be reversed via fine-tuning (Gade et al. 2023; Lermen et al. 2023) This implies that safety audits of models before fine-tuning are likely to be necessary for models which will be made available to fine-tune.

Not only is it important to have process audits to ensure the end product is safe, but there are also a number of mitigations for bias and similar problems that can be applied throughout the model training process (Bellamy et al. 2018). For these and other reasons, we expect defense in depth and lifecycle threat analysis to be very valuable (Ee et al. 2023). If it proves useful for external auditors to audit AI systems over multiple steps during training, and throughout the process, then they will need in-depth engagement throughout

the development process. This is explored further by Campos et al. (2025)

2.4 Rapid progress requires evolving standards

In other domains, best practices for risk evaluation have evolved to recognize the need for proactive or dynamic risk evaluations (Rasmussen and Suedung 2000). This means that in fields much less dynamic than AI, proactive engagement with risk management is vital. In artificial intelligence, the dizzying pace of advances makes adaptability even more critical.

However, changing rules can easily create incoherent or conflicting standards, and different standards have already created a difficult to follow patchwork of best practices, requirements, and an emerging regulatory tangle of overlapping and burdensome requirements. There is currently no group that can decide when or if obsolete proposals should be ignored, or when regulations have been superseded by practice. Because standards bodies do not exist, and a wide variety of academic work suggests approaches exist alongside unclear regulatory standards, and model developer pronouncements about standards they abide by, current best practices and requirements will remain unclear.

2.5 Continued monitoring and ongoing analysis

Monitoring is not only needed throughout the development lifecycle, but also post-deployment. This is already occurring in places, for example OpenAI's preparedness framework (OpenAI 2023c), where the model provider monitors usage to identify abuse or misuse. This process, however, is not currently done in a principled or public way—there are no clear commitments to report on these ongoing 44 evaluations, much less precommitments for specifics of what to report. There is also no architecture for bug bounties or patch notes. The current process is also reactive, whereas audits based on predetermined standards are proactive. Ideally, a principled and public process that tries to anticipate future misuse would include ongoing monitoring, rather than post-hoc reactions when capabilities are discovered.

Additionally, the OpenAI framework emphasizes forecasting future risks to develop adequate safety and security measures ahead of time. For example, they intend to privately fine-tune models towards dangerous applications to determine how far their dangerous capabilities can be stretched, and continuously monitor their deployed models in worst case settings. This framework's inclusion of continuously monitoring capabilities along various dimensions could address many of the requirements we have outlined for AI auditing—though this auditing is done in-house, with no reporting requirements, and no longer term commitment to the process, so it obviously fails the goal of providing public or regulatory assurance.

2.6 Inadequate audits can cause failures

Major AI developers expect that work at the AI frontier will increasingly focus on designing AI systems to autonomously pursue large-scale goals (Anthropic, OpenAI 2023b; Christiano 2018). As such systems are developed, then audits must also ensure both that these autonomous agents do not have discriminatory or otherwise unethical consequences, and that their goals are and remain more broadly aligned with human values and ethics over time. However, internal teams repeatedly testing these sophisticated AIs against simple audits could influence them to develop unintended and dangerously deceptive behaviors. This could directly cause dangerous failures, or at least lead to complacency about safety as dangers are merely papered over by trained success in audits.

Christiano identifies three specific dynamics that make it hard to detect misalignment in AI systems. First, the "simple training game" is, where an AI behaves sycophantically to make deception more difficult to notice. Second, "deceptive alignment" is where AI has hidden goals other than what is apparent in its performance and predicts what will fulfill these goals over a long timescale. Both have already been seen to some extent in the wild (Park et al. 2024). Third, "gradient hacking" is where AI prevents training from changing its hidden long-term goals (Christiano 2018). It is unclear the extent to which any of these will occur, but there is at least a strong reason to be concerned that poorly managed audits will create these risks. Of course, auditing versions of models without alignment or safety mitigations implemented—as suggested in the OpenAI Preparedness Framework (OpenAI 2023c)—is one crucial way of addressing this risk, as is ensuring that audits are not part of the internal development process.

3 Governance for safety

Building on the work of Raji et al. (2022a), we note that culture can be studied at multiple levels. The nuclear industry (IAEA 2013) draws on Schein (1992) to look at three levels of culture, artifacts, as discussed by Raji, espoused beliefs and values, and underlying assumptions. As the nuclear industry has found, technical safeguards are insufficient, and building a culture of safety is critical in minimizing failures. The airline industry, among others, has found the same (CANSO 2008). A number of concrete suggestions for how this impacts audits more specifically will be highlighted below, but the overarching theme is that change in the industry requires industry buy-in to standards and norms, of the

type facilitated by standards boards in which they have an active role, but undermined by proliferating standards.

However, as pointed out by Manheim (2023a) there are many challenges in building a culture of safety in the AI industry. In the context of auditing and governance, these challenges are critical (Robinson et al. 2025). There are two aspects to addressing safety culture that are relevant to audits; the way that safety culture makes audits more effective, and the way that audits can ensure safety culture. For both, formulaic and fixed standards seem insufficient. Until recently, the AI industry was not at all focused on these issues, and it remains complacent about safety.

3.1 Safety culture enables effective audits

Auditing systems is not, in and of itself, a method to reduce risks. Instead, it functions as part of an ecosystem of regulation, risk analysis, internal controls, public oversight, and changes to the systems themselves in response to all of these factors. Because of this, the value of auditing in improving safety lies partly in how those being audited view the process and anticipate or respond. In domains where audit is adversarial or viewed as superfluous, knowledge of problems does not lead to addressing them; if any attention is paid, it can lead to deception in order to move forward, or at best patches that minimally address the deficiencies. Designing metrics that can be evaluated in such an environment is possible, but very challenging - and a key strategy for overcoming the challenge is adaptability and flexibility (Manheim 2023b). It seems likely that standards boards would make this type of adaptability and flexibility easier.

In contrast to such antagonistic relationships between auditing and business, there are many domains where auditing is aligned with the mission and values of those audited. In safety-critical industries like the nuclear industry or aviation, safety of the power plants or airplanes is vital to the overall success of an organization, and safety failures are business failures. These are industries where safety is valued, and firms take note of highlighted deficiencies, which can then be integrated into plans for improvement. If it is important for the AI industry to emulate the positive examples, a critical role of those tasked with monitoring and evaluating safety of AI systems is promoting cultural change in the industry. Unless and until safety is a goal of those producing the models, among other things, auditing will lead to adversarial engagement, or at best grudging compliance. Such cultural change also benefits greatly from collaborative engagement in standards settings.

3.2 Internal auditing is insufficient

Internal auditing is one way to try to build alignment between auditing and development. This has a number of advantages over external audits, as clearly discussed by Raji et al. (2020) However, public assurance requires more than just internal auditing. Audits—whether internal, external, or ideally both—must be guided by independent reporting standards and have lines of reporting outside of the company structure to ensure credibility (Schuett 2024). Among other concerns, this is because internal auditing without broad input about the concerns that must be addressed can easily become safety-washing (Lazar and Nelson 2023). A business that declined to have their financial records audited externally because they did sufficient checking internally would be farcical, and a business which decided to set the standards which it used would be similarly unacceptable.

For this reason, attempts to bring auditing under the control of firms developing models, such as OpenAI's push to keep their red-teaming under their own control (OpenAI 2023a), is a worrying development. Similarly, the "Frontier Model Forum" (OpenAI 2023b), while laudable as an attempt to ensure that standards exist, could allow industry to maintain control over what those standards will be.

At the same time, developers' concerns should also be a critical consideration. Developers claim that proprietary concerns merit keeping capabilities evaluations in house, and some go so far as to suggest that these concerns merit exclusion of third-party evaluators (Weidinger et al. 2023). As noted, this is fundamentally opposed to trusted audits, but the risk is critical. Standards setting bodies must find ways to balance external inputs and requirements with internal control, as has been done by standards boards and regulators in other industries.

For example, model auditing processes, and the standards that inform them, must be careful to take stringent steps to protect developers' intellectual property. Toward these ends, it is not unreasonable for third party evaluators to be embedded in organizations and subject to confidentiality agreements—it can work to the advantage of auditors to have unfettered access to engineers and scientists developing the systems, and without such access, evaluations might be far less effective. Despite that, contrary to the status quo, the confidentiality agreements must themselves be public, to allow consumers of these independent audits to know what the auditors may be prevented from saying. Negotiations between model auditing providers and firms, however, make it difficult for auditors to insist on such terms.

Only external standards that balance these and similar considerations, developed openly as part of an independent and iterative process with industry participation, have been shown in other comparable cases to lead to reasonable assurances of safety. In such a broad-based process, those who are not dependent on firms, including regulators and the public, must push for more openness, where auditors would not be willing or able to do so. A forum which sets standards is an ideal setting for this.

3.3 Audits should provide transparency, not create requirements

One critical component of an audit is to create common knowledge of what is and is not true, and what has or has not been done. While it is important for standards for audits to be created, that does not mean an audit itself implements or enforces a standard. For this reason, an audit may show that a system is capable of a certain class of abuse, or is not, and in either case, that is an acceptable audit outcome, even when it is a substantive failure. Similarly, a firm which is audited financially and is found to have debts in excess of assets may be insolvent, but the audit was not failed. The equivalent determination for an AI system or company would be that it abides by a given standard or follows a regulatory requirement, or not.

For example, there are debates about the legality or morality of training models on copyrighted material. While an audit standard will not resolve this debate, an audit can determine whether such material was included in the training data, what efforts were made to exclude both copyrighted material, and material which creators explicitly request to have excluded. It will then be up to regulators and the public to insist on reasonable rules. Similarly, audits of AI models should detail whether internal or external red teaming processes occurred, and what they did or did not cover. This is different than a requirement to do so—and whether such a requirement should exist is again, up to regulators and lawmakers, not those who design or carry out audits.

3.4 Audits must allow for failure

Despite the above points, if an audit cannot be failed, it cannot serve its full purpose as an audit. In domains where auditing does function as an independent check on firms, such as in financial accounting, the auditors will collect instances of material misrepresentation and report them, rather than simply demanding compliance. This type of review is mandated by the US Financial Accounting Standards Board (FASB), and the international equivalent, the International Accounting Standards Board (IASB). The requirement to report misrepresentation rather than quietly amend them is because post-hoc compliance is ineffective at addressing core concerns; a firm which lies to auditors should not have their remaining claims trusted. For this reason, requiring them to correct the specific instances in which the misstatements occurred does not change the need to report that misrepresentation occurred.

Similarly, a firm which builds unsafe systems for audit, expecting that the unsafe system will be remedied during the audit, has already failed to build a safe system. For this reason, allowing them to correct the specific instances in which the model is dangerous is insufficient. Not only is it unlikely to fully address the problems uncovered, but the safety failures should be reported by auditors even when the firm addressed the failures before release. This means that in practice, auditors should expect to see the reports for safety failures that were caught internally by a responsible company's monitoring and reporting, not by auditing - and the absence of such events being flagged before auditing should be seen as a red flag, not a positive sign.

Unfortunately, a variation on this can easily be used perversely to render auditing far less valuable. Audits that are within the development cycle are often used as a way to iterate on failure rather than as a check. In such cases, the supposed existence of an audit is (perhaps unintentionally) actually enabling development rather than acting as a check. This is especially acute if the audit is under the control of those being audited, either because the audit is performed internally, or with nondisclosure agreements. In fact, this is explicitly the process that OpenAI's Preparedness framework adopts, where high or critical risk systems are not stopped, and instead, mitigations are put in place so that the systems can be argued to be below the internally defined threshold (OpenAI 2023c). For this reason, audits should be independent of development, and should have the mandate to honestly report what was found, acknowledging failures rather than ensuring that failures are individually addressed.

Even for less serious concerns, and for audit requirements that do not impose a standard at all, as discussed above, allowing firms to address problems without announcing the issues or changes prompted by the audit is a failure. For example, models that are fine-tuned to eliminate individual problems still have capabilities that can be abused by prompt engineering, or by fine-tuning to reverse the new safety measures. Unfortunately, quietly and iteratively fine-tuning failures away as they are discovered, either before or after release, seems to be the current modus operandi (Mitchell et al. 2019). But post-hoc patches do not provide public knowledge of fallibility and cannot serve the same purpose as auditing. For parallel reasons, audits of training data that identify biases, or problematic content of various types, must report this fact even if the training data is amended to exclude that content.

Non-public failures and responses also fail to alert others about measures which individual firms found to be effective. This is true of a broad variety of harm mitigation measures, and while businesses developing AI have reasonable incentives not to publicize proprietary methods, if the culture of proprietary technology applies to safety and ethics, insights about how to address problems will remain unknown, and safety of other systems will be impeded. In this case, the benefits of openness parallel early claims for the necessity of open-source in machine learning endorsed by both Bengio and LeCun (Sonnenburg et al. 2007), rather than their recent debates over whether model weights should be open-source (Nuñez 2023).

At the same time, auditors will presumably have the prerogative not to publicize measures which would be compromised by public knowledge.

3.5 Auditing for governance and culture

To conclude the discussion of the relationship between auditing and building a safety culture, we reiterate the point that a sufficient audit must have a scope that includes the firm and its governance, management, and culture. (Robinson et al. 2025) Ideally, auditing should inform everything from strategic goals to monitoring. Filyppova et al. (2019) Once that is the case, auditing should also be used to understand and monitor the safety culture.

For example, audits should include a review of risk management systems internal to the firms being audited. A company with a healthy safety culture would have well documented safeguards for internal reports of safety failures and incentives for reporting risks and failures, would use risk registers to document what risks they address and which have been raised, would have robust systems in place to address each of the risks, would have systems for raising and addressing internal safety concerns, and would have clear procedures for stopping development or training if concerns are raised. All of these systems would have clear processes in which actions taken are documented, and in each case, an audit would create artifacts, as discussed by Raji et al. (2020) documenting how the systems have been used, the concerns and risks they currently address, and exercises done to test the system's ability to respond.

Surveys of employees' view of risks and their understanding of how the company engages in proactive identification of new sources of risk are another type of tool that auditors can and should use. Beyond this, however, firms can also use similar tools internally to proactively assess and reinforce safety culture. Similarly, process improvement targeted at improving risk management can be a result of audit recommendations, but far better than needing auditor feedback is auditors ensuring that the firm itself has internal process improvement methods that focus on safety and identify risks and concerns, and address them openly, in ways that auditors can observe. To ensure this, it seems likely that auditors should, as a part of their audit, ensure that management and oversight systems are in place for reporting failures, either publicly, or at least to the auditors. All of these suggestions are issues that an audit standards board should discuss and choose how auditors should implement.

4 Auditing standards body, not standard audits

We do not have sufficient standards for software safety in general, nor for ethical usage of such systems, compounding the problems with making AI systems safe. Building such standards is important, but disparate proposals and a lack of unification of standards undermines accountability. Industry standards are a useful step, but one-time processes to develop standards has turned into ad-hoc changes to standards with each new model. As noted initially, in comparable industries, there are bodies explicitly tasked with ongoing development of standards, and because of the likely importance of AI in industry in coming years, no less should be expected here.

One key question is the scope of such standards, and this is another area where an audit standards board would be useful, as we discuss below.

4.1 Boards providing guidance, not static rules or regulations

Best practices from other fields are that standards develop as risks and practices change, and this process is ongoing. Not only are standards constantly evolving, technical research informs audit boards like FASB for accounting, and regulators like the US Occupational Safety and Health Administration (OSHA) for safety. And in such cases, the standards are part of an ecosystem, not an endpoint. This is uncontroversial; FASB is asked to provide guidance on a routine basis whenever new accounting issues arise (Tucker 2002) and OSHA provides interpretation letters to clarify how requirements apply in new cases (Pepper 2001).

We have also seen this fail in other domains. Organizations like NIST (Schooley 2000) provide important standards for domains like cryptography, updated routinely, and provide technical reports on blockchain systems, but regulation and standards cannot keep up with the pace of new applications and risks for digital securities. The faster the systems develop and change, the more agile the standards must be. Flexibility is therefore critical, which bolsters the case for independent boards rather than direct government regulation. It also highlights the importance of guidelines for auditors that can be interpreted by external standards groups into specific audits and methods, rather than government agencies attempting to constantly change requirements.

The likely alternative to updated standards can cause perverse effects. Specific requirements and maximums imposed by regulation often become a target when they

relate to a system property of interest, per Shorrock's Law of Limits (Shorrock 2019). In the case of AI, a requirement that systems not score over some limit on a risk scale, or not have a racial or other bias that exceeds some numerical value, would require checking of compliance, and if there is any expense associated with compliance, this produces implicit pressure to get as close as possible to that limit without exceeding it.

In parallel cases, government regulators have required corporations to follow guidelines developed privately. For example, ISO standards, which are developed by experts from a variety of national standards bodies, business groups, and others, are not themselves legal standards. Even when not mandated, standards are often stipulated in contracts. This can create a norm of compliance with an evolving standard, and regulators often later adapt or adopt the standards.

For systems that are under continual development, which are retrained or fine-tuned on recent public data, any public knowledge about the contents of a test could compromise the integrity of the test.

For this reason, standards development needs to occur in parallel to, and must be informed by (but separate from), model development. That is, auditing should not be an internal process to enable development, but instead be a check on the model developers, as discussed below. Therefore, standards setting groups should define tests which must be passed by models both initially, and if and when new concerns or capabilities emerge—and otherwise halt development or release of a model.

4.2 Balancing principles and rules

Whichever class of oversight and development of audit standards occurs, one critical question is whether to use principles based standards, or rule based standards. This parallels but does not exactly match the question of static versus adaptable standards; there can be unhelpfully static principle based standards, and dynamic and rapidly adapted rule based standards, but these are harder than having good principles that give rise to adaptability. Further, principles based standards have been shown to lead to better auditing than rule base standards in other important ways (Peytcheva et al. 2014; Schipper 2003).

Principle based standards also provide much more adaptability for rulemakers to provide useful guidance. For example, in the above discussion differentiating between requirements and standards, it was noted that transparency about practices can be a viable alternative to mandating specific practices. In this case, transparency should be a principle, not a specific requirement, for example, a requirement to announce that some specific method was (or was not) used. Without such a principle, firms must fulfill requirements narrowly and report compliance, making the audits less flexible and more of a routine. The alternative provides flexibility to do more than other firms and report that, or to use newer and more powerful methods for assuring some specific objective is achieved - while still ensuring that auditors and consumers of the audit can understand if the methods used fall short in some regard.

But as is often the case for dichotomies, some combination is both inevitable and wise. Principles relevant in this context could include transparency about model training and inputs, preregistration of training details including safety guards, commitments to which safety measures are employed during each phase of release, commitments for monitoring and mitigating negative impacts, and external verification of the above. Rules and minimum standards implementing the principles will be needed, but these can be adaptive guidelines, as static rules to accomplish the same ends will inevitably be both incomplete and restrictive, and require adaptation in individual cases.

4.3 Scope and applications-specific audits

Defining the scope of AI audit standards is both complex and critical, as AI spans narrow machine learning models, general AI systems, and potentially future fully general intelligences. While domain-specific regulations provide oversight in fields like medicine and law enforcement, these typically assume human intent and expert judgment, leaving gaps in AI accountability. One current approach seems to be that AI systems are either narrow and audited in individual domains, or are general but are trained to refuse to answer within domains where issues exist. Refusing or qualifying responses in restricted domains is a temporary partial safeguard, but among other shortcomings it fails to address what the EU AI act refers to as systemic risks.

A robust incentive structure for AI auditing must ensure not only useful audits of general and frontier models but also rigorous, risk-specific evaluations, including application-specific audits than apply to narrower models and applications. Different risk classes require distinct expertise, making red teaming and audit specialization essential (Frase 2023). The lack of AI-specific expertise in many of the relevant application areas further complicates both writing or adapting relevant regulations, and enforcement, underscoring the benefits of a dedicated AI Audit Standards Board to help guide and adapt auditing frameworks to different industries.

High-risk applications, such as law enforcement and medicine, will require additional work to ensure that audits consider the domain-specific risks and requirements. While regulatory bodies govern human professionals in these fields, they are sometimes ill-equipped to evaluate AI-driven decision-making.

Therefore, a multi-layered audit approach is likely useful for addressing this, covering both general AI systems and application-specific risks. For instance, the AI Audit Standards Board could develop adaptable auditing protocols to evaluate (1) System-level audits, ensuring AI models meet baseline ethical, safety, and transparency criteria, (2) governance level audits, as explained above, adapted to match governance in specific industries, and (3) application audits, setting additional standards based on the model's intended or actual use.

On this third point, for example, a generative AI model creating CAD files for medical applications like dental implants should be held to higher safety and liability standards than one generating content for car parts, and that needs more oversight than one used to generate content for online games. Similarly, models used in medical diagnostics or legal analysis should be subject to stringent ethical and performance criteria, not just general AI safety benchmarks.

AI's increasing generality poses further challenges. User disclaimers stating that AI should not be relied upon for legal or medical advice are legally ambiguous and ethically insufficient, as users still act on AI-generated recommendations. Therefore, to the extent that a model can be urged to provide a service, there is an ethical if not legal requirement to ensure it does so responsibly—so that general-purpose AI systems could be subject to many relevant domain-specific standards, if they are capable of high-risk functions.

Given AI's rapid evolution, as discussed earlier, a fragmented, static audit approach is insufficient. Instead, coordinated oversight among regulators, industry leaders, auditors, ethicists, and the public is necessary. The AI Audit Standards Board would ideally used holistic approaches and expert groups able to respond dynamically to changing needs and define the scope of audits needed, rather than evaluating and addressing each class of harm with different sets of incomplete and overlapping audits. In the best case, it would serve as a central forum to harmonize global audit standards, enable ongoing risk assessments, and provide guidance on emerging challenges, ensuring that AI governance remains adaptive, transparent, and enforceable.

4.4 Safety-washing and industry dynamics

Despite AI's unique challenges, it shares regulatory failure modes with other industries.

Unfortunately, many sectors suffer from regulatory capture, competition driving minimal compliance, and audits that prioritize often unsuccessful attempts at preserving reputation over substance. This pattern is evident in industries from chocolate production and modern slavery (Gutierrez-Huerter et al. 2023) to mining (Gold et al. 2015; Nygren 2018), and seems likely to plague AI safety (Guest et al. 2023) and AI ethics (van Maanen 2022).

Current standardization efforts risk being shaped by corporate interests, undermining their effectiveness. Unfortunately, competitive pressure has led to abandoning cautious approaches like the staged release of GPT-2 (Shevlane 2022). Algorithmic decision-making has already caused significant harm (O'Neil 2016), and as AI adoption accelerates in high-stakes domains, industry capture would be disastrous. While initiatives like the EU AI Act attempt to integrate corporate input, any auditing body must coordinate with regulators and standards organizations to maintain independence and credibility.

Regulatory challenges are further exacerbated by profit-driven urgency, which, paradoxically, threatens long-term industry stability. Public backlash against AI failures-justified or not-is inevitable if standards are dictated by narrow interests and are untrusted, or unenforced. Recent events illustrate the problem, such as companies not releasing full model cards for reasoning models until well after deployment. Thus, industry should support proactive, coordinated auditing efforts and clear rules to prevent reactionary regulation that may be harsher and less effective.

Historically, industries have succeeded or failed based on their commitment to strong ethical and safety standards. Given AI's transformative role, it is in the sector's self-interest to establish robust, independent, and transparent auditing practices. Ensuring redundancy, rigorous certification, and audit consistency is critical, especially for frontier models with unprecedented risks.

Unfortunately, instead of fostering consensus, some companies are preempting regulation through voluntary frameworks that fail to provide genuine oversight. Efforts like Microsoft's Responsible AI Maturity Models (Vorvoreanu et al. 2023) and industry-led groups such as Responsible.AI Responsible.AI (2014) are clear attempts to sidestep regulatory scrutiny. Maturity models, for example, are often useful but are used primarily in otherwise unregulated spaces. This offers the illusion of responsibility while excluding regulators and public oversight. Similarly, OpenAI's red teaming network undermines independent red-teaming and transparency (OpenAI 2023a). Narrow, non-participatory efforts ultimately undermine audits and, ironically, work against the long-term interests of AI companies themselves.

4.5 Implementation concerns

A variety of reasonable concerns exist about how such a board can be created and run, and the details of implementation are critical. A variety of challenges exist to the approach outlined, including jurisdictional issues and the need for coordination both internationally (ÓhÉigeartaigh et al. 2020) and, as discussed above, between companies and other stakeholders.

A board would ideally work to build auditing standards alongside and incorporating, rather than competing with, regulators and international agencies (Gruetzemacher et al. 2023). This might fail, but existing efforts, such as the EU AI Act and the AI Safety Institute network members, would both benefit from a clear public coordination mechanism not tied to a single government or law. Ideally, this would greatly reduce, rather than increase, bureaucracy compared to the likely alternatives - making this among the best outcomes for industry as well.

If pursued, the exact structure of a board, its funding model, and its governance structure will all be critical. For these reasons, the groups which need to push for such a standards body are not the academics studying the problem, or papers like this one dictating answers, but discussion among a coalition that includes companies building these systems, auditors performing evaluations, and regulators, government bodies, and international agencies who are charged with oversight. With buy-in from these groups, and concerted effort, similar efforts have succeeded in the past, and learning from those efforts will be critical moving forward. The groups can then decide whether to just publish standards, or should accredit auditors and credential individuals, and choose to directly produce or oversee certain classes of test, or just serve as a clearinghouse for audit results and analyses.

5 Conclusion

An audit standards board would address many issues with AI audits and the ecosystem generally, and efforts in this direction seem valuable. As noted initially, standards boards for safety and similar audits are common across many industries, in various forms. And despite differences between AI and those other fields, the justification is the same; audits should be fair, trusted, and fit to the purpose of enabling stakeholders to make determinations about the facts and addressing risks. Those goals cannot be accomplished without having the standards set independently, collectively, and transparently.

The alternative current proliferation of proposed standards and diverse and overlapping regulation seems unlikely to accomplish this goal, to the detriment of all involved, whereas a board of people actively building auditing systems and assisting those performing audits, with transparent processes and cooperation from regulators and industry, seems far more promising.

Such a board would not, in and of itself, solve problems with regulatory capture, much less with the fundamentally difficult problems of building artificial intelligence that does not exacerbate current harms or create larger future risks. Open work by such a board would, however, make it clearer to everyone what progress was or was not occurring. It would also allow current and future safety, audit, and governance research to have a clear audience and adoption pathway. To the extent that AI governance is an information problem (Anderljung et al. 2023a), a standards board can help remediate that problem. And to the extent that ethical and safe AI can be enhanced by wider sharing of best practices, standards, and methods, this seems like the right direction.

Consensus about AI audit as one important path forward is a promising recent development, and we are hopeful that this next step, of having an active standards board that sets audit standards and provides resources for performing them, also becomes a widely-accepted next step in building more fair, safer and ultimately broadly beneficial AI systems.

Acknowledgements We would like to thank two of the anonymous reviewers, and thank the broader AI safety and AI ethics communities for the discussions leading to this work

Author contributions DM wrote the initial draft with assistance and conceptual contributions from MS. SM, MB, and RG then contributed significant ideas and revisions to the manuscript and helped revise and polish it. SM designed the diagram.

Funding Open access funding provided by Technion - Israel Institute of Technology.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Anderljung M, Barnhart J, Leung J, Korinek A, O'Keefe C, Whittlestone J, Wolf K (2023a) Frontier AI regulation: managing emerging risks to public safety. arXiv preprint arXiv:2307.03718

Anderljung M, Smith ET, O'Brien J, Soder L, Bucknall B, Bluemke E, Chowdhury R (2023b) Towards publicly accountable frontier LLMS: building an external scrutiny ecosystem under the aspire framework. arXiv preprint arXiv:2311.14711

Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, Ahn J (2023) Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805

Anthropic (2023) Anthropic's responsible scaling policy
Avin S (2023) Frontier AI regulation blueprint. Center for the Study of Existential Risk (available online)
Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Zhang Y (2018) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943
Bhaumik S, Rana S, Karimkhani C, Welch V, Armstrong R, Pottie K, Amico RD (2015) Ethics and equity in research priority-setting: stakeholder engagement and the needs of disadvantaged groups. Indian J Med Ethics 12(2):110–3
Biden J (2023) Eexecutive order 4110. Technical report, United States White House
Biene-Hershey M (2007) It security and it auditing between 1960 and 2000. The history of information security. Elsevier Science BV, Amsterdam, pp 655–680
Birhane A, Kalluri P, Card D, Agnew W, Dotan R, Bao M (2024) AI auditing: The broken bus on the road to AI accountability. arXiv preprint arXiv:2401.14462. https://arxiv.org/abs/2401.14462
Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258
Campos S, Papadatos H, Roger F, Touzet C, Quarks O, Murray M (2025) Bridging the gap between current AI practices and established risk management. A frontier AI risk management framework
Carpenter D (2013) Corrosive capture? The dueling forces of autonomy and industry influence in FDA pharmaceutical regulation. In: Preventing regulatory capture: social influence and how to limit it. pp 2228–2232
Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, Hadfield-Menell D (2023) Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217
Christiano P (2018) Three levels of threat obfuscation
Civil Air Navigation Services Organisation (CANSO) (2008) Safety Culture Definition and Enhancement Process. https://www.icao.int/NACC/Documents/Meetings/2018/ASBU18/OD-10-Safety%20Culture%20Definition%20and%20Enhancement%20Process.pdf
Commission L'intelligence Artificielle (2024) Ia: Notre ambition pour la france
Corso A, Karamadian D, Valentin R, Cooper M, Kochenderfer MJ (2023) A holistic assessment of the reliability of machine learning systems. arXiv preprint arXiv:2307.10586
Courville S, Parker C, Watchirs H, Costanza-Chock S, Raji ID, Buolamwini J (2003) Introduction: auditing in regulatory perspective. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, vol 25, no 3, pp 179
Department for Science Innovation & Technology DSIT (2023) Introducing the AI safety institute
Ee S, O'Brien J, Williams Z, El-Dakhakhni A, Aird M, Lintz A (2023) Adapting cybersecurity frameworks to manage frontier AI risks: a defense-in-depth approach
Faveri B, Johnson-León M, Sylvester P, Hui Kyong CW, Hannák A, Mendoza M, Broussard M, Enikolopov R, Nelson A, Sandvig C, Sesan G, Sporle A, Srihari R, Srinivasan J, Wilson C, Zou M (2025) Towards a global AI auditing framework: assessment and recommendations. Synthesis report 2024.3, International Panel on the Information Environment (IPIE)
Filyppova S, Kholod B, Prodanova L, Ivanchenkova L, Ivanchenkov V, Bashynska I (2019) Risk management through systematization: risk management culture
Frase H (2023) One size does not fit all: assessment, safety, and trust for the diverse range of AI products, tools, services, and resources. Georgetown University, Center for Security and Emerging Technology
Gade P, Lermen S, Rogers-Smith C, Ladish J (2023) BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. arXiv preprint arXiv:2311.00117
Gold S, Trautrims A, Trodd Z (2015) Modern slavery challenges to supply chain management. Supply Chain Manag 20(5):485–494
Gruetzemacher R, Chan A, Frazier K, Manning C, Los Š, Fox J, Hernández-Orallo J (2023) An international consortium for evaluations of societal-scale risks from advanced AI. arXiv preprint arXiv:2310.14455
Guest O, Aird M, ÓhÉigeartaigh S (2023) Safeguarding the safeguards: how best to promote AI alignment in the public interest. arXiv preprint arXiv:2312.08039
Gutierrez-Huerter O, Gold GS, Trautrims A (2023) Change in rhetoric but not in action? Framing of the ethical issue of modern slavery in a UK sector at high risk of labor exploitation. J Bus Ethics 182(1):35–58
Hall JA, Hazell J (2015) Cengage Learning. Information technology auditing. Oxford Internet Institute, University of Oxford, Centre for the Governance of AI
Hammond L, Chan A, Clifton J, Hoelscher-Obermaier J, Khan A, McLean E, Smith C, Barfuss W, Foerster J, Gavenčiak T, Han TA, Hughes E, Kovařík V, Kulveit J, Leibo JZ, Oesterheld C, de Witt CS, Shah N, Wellman M, Bova P, Cimpeanu T, Ezell C, Feuillade-Montixi Q, Franklin M, Kran E, Krawczuk I, Lamparth M, Lauffer N, Meinke A, Motwani S, Reuel A, Conitzer V, Dennis M, Gabriel I, Gleave A, Hadfield G, Haghtalab N, Kasirzadeh A, Krier S, Larson K, Lehman J, Parkes DC, Piliouras G, Rahwan I (2025) Multi-agent risks from advanced AI
Hazell J (2023) Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. Montreal AI Ethics Institute. https://montrealethics.ai/large-language-models-can-be-used-to-effectively-scale-spear-phishingcampaigns/
Hendrycks D (2025) Introduction to AI safety, ethics, and society. Taylor & Francis, Milton Park
Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, Perez E (2024) Sleeper agents: training deceptive LLMs that persist through safety training. arXiv preprint arXiv:2401.05566
IAEA (2013) Regulatory oversight of safety culture in nuclear installations. International Atomic Energy Agency (IAEA) (1707)
Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov from Data (TKDD) 6(4):1–21
Keyser L (1974) Apollo experience report. The role of flight mission rules. Lyndon B. Johnson Space Center
Khalid F, Hanif MA, Rehman S, Ahmed R, Shafique M (2019) TrISec: training data-unaware imperceptible security attacks on deep neural networks. In: 2019 IEEE 25th international symposium on online testing and robust system design (IOLTS), Rhodes, Greece, pp 188-193
Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, Lomas E (2021) Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms
Krijger J, Thuis T, Ruiter M (2023) The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations. AI Ethics 3:355–367
Lawrence XY, Woodcock J (2015) FDA pharmaceutical quality oversight. Int J Pharm 491(1–2):2–7
Lazar S, Nelson A (2023) AI safety on whose terms? Science 381(6654):138–138
Lermen S, Rogers-Smith C, Ladish J (2023) Lora fine-tuning efficiently undoes safety training in llama 2-Chat 70B. arXiv preprint arXiv:2310.20624

Liang P (2023) Holistic evaluation of language models, 2022. Published in Transactions on Machine Learning Research (TMLR). arXiv:2211.09110
Lombe M, Sherraden M (2013) Inclusion in the policy process: an agenda for participation of the marginalized. New horizons for policy practice. Routledge, London, pp 109–123
Maanen G (2022) AI ethics, ethics washing, and the need to politicize data ethics. DISO 1:9
Manheim D (2023) Building a culture of safety for AI: perspectives and challenges
Manheim D (2023) Building less-flawed metrics: understanding and creating better measurement and incentive systems. Patterns 4(10):100842
Meemi (2025) Meemi's shortform: Frontiermath was funded by openai. Lesswrong Personal Blog Post, January. Accessed 12 Mar 2025
Miceli M, Posada J (2022) The data-production dispositif. Proc ACM Hum Comput Interact 6(CSCW2):1–37
Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Gebru T (2019) Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency, pp 220-229
Mökander J, Schuett J, Kirk HR, Floridi L (2023) Auditing large language models: a three-layered approach. AI Ethics, 1–31
Moscou K, Kohler JC (2018) Pharmacogovernance: advancing pharmacovigilance and patient safety. Social and administrative aspects of pharmacy in low-and middle-income countries. Academic Press, Cambridge, pp 403–418
Mouton CA, Lucas C, Guest E (2023) The operational risks of AI in large-scale biological attacks: a red-team approach. RAND Corporation, Santa Monica
NIST (2023) AI risk management framework: AI RMF (1.0). Technical report NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg, MD
Nuñez M (2023) AI pioneers yann lecun and yoshua bengio clash in an intense online debate over AI safety and governance
Nygren M (2018) Safety management on multi-employer worksites: responsibilities and power relations in the mining industry
ÓhÉigeartaigh SS, Whittlestone J, Liu Y, Zeng Y, Liu Z (2020) Overcoming barriers to cross-cultural cooperation in AI ethics and governance. Philos Technol 33:571–593
O'Neil C (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group. https://weaponsofmathdestructionbook.com/
OpenAI (2023a) OpenAI red teaming network
OpenAI (2023b) Frontier model forum. A joint announcement from Anthropic, OpenAI. Microsoft, and Google
OpenAI (2023c) Preparedness framework beta. December, 18
OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo A, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, de Avila Belbute Peres F, Petrov M, de Oliveira Pinto HP, Pokorny M, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B (2024) Gpt-4 technical report
Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D (2024) AI deception: A survey of examples, risks, and potential solutions. Patterns. 5(5):100988. https://doi.org/10.1016/j.patter.2024.100988
Pepper TG (2001) Understanding osha: a look at the agency's complex legal & political environment. Prof Saf 46(2):14
Peytcheva M, Wright AM, Majoor B (2014) The impact of principles-based versus rules-based accounting standards on auditors' motivations and evidence demands. Behav Res Account 26(2):51–72
Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 33–44. https://doi.org/10.1145/3351095.3372873
Raji ID, Kumar IE, Horowitz A, Selbst A (2022a) The fallacy of AI functionality. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 959-972
Raji ID, Xu P, Honigsberg C, Ho D, D I, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Barnes P (2022b) Outsider oversight: designing a third party audit ecosystem for AI governance. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, pp 557-571
Rasmussen J, Suedung I (2000) Proactive risk management in a dynamic society. Swedish Rescue Services Agency
Responsible.AI (2024) (Responsible AI Institute). Our responsible AI maturity model
Robb L, Candy T, Deane F (2023) Regulatory overlap: a systematic quantitative literature review. Regul Govern 17:1131–1151
Robinson B, Murray M, Ginns J, Krzeminska M (2025) Why frontier AI safety frameworks need to include risk governance. Technical report, The Centre for Long Term Resilience
Sandbrink JB, Ashurst C, Karpathy A, Mishra S (2023) Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools. arXiv preprint arXiv:2306.13952. https://arxiv.org/abs/2306.13952

Schein EH (1992) Organizational culture and leadership. Jossey-Bass Publishers, San Francisco
Schipper K (2003) Principles-based accounting standards. Account Horiz 17(1):61–72
Schooley JF (2000) Responding to National Needs: The National Bureau of Standards Becomes the National Institute of Standards and Technology: 1969–1993, vol 955. National Institute of Standards and Technology, Technology Administration
Schuett J (2024) Frontier AI developers need an internal audit function. Risk Anal. https://doi.org/10.1111/risa.17665
Shah R, Pour S, Tagade A, Casper S, Rando J (2023) Scalable and transferable black-box jailbreaks for language models via persona modulation
Sharkey L, Ní Ghuidhir C, Braun D, Scheurer J, Balesni M, Bushnaq L, Stix C, Hobbhahn M (2023) A causal framework for AI regulation and auditing. Apollo Research
Shevlane T (2022) The artefacts of intelligence: governing scientists' contribution to AI proliferation
Shevlane T, Farquhar S, Garfinkel B, Phuong M, Whittlestone J, Leung J, Kokotajlo D, Marchal N, Anderljung M, Kolt N, Ho L (2023) Model evaluation for extreme risks
Shin D, Shin EY (2023) Human-centered AI: a framework for green and sustainable AI. Computer 56(6):16–25
Shorrock S (2019) Shorrock's law of limits
Sonnenburg S, Braun ML, Ong CS, Bengio S, Bottou L, Holmes G (2007) The need for open source software in machine learning. J Mach Learn Res 8:2443–2466
Tidjon LN, Khomh F (2023) The different faces of AI ethics across the world: a principle-to-practice gap analysis. IEEE Trans Artif Intell 4(4):820–839
Tucker TN (2002) It really is just trying to help: the history of FASB and its role in modern accounting practices. NCJ Int'l L Com Reg. 28:1023
Vorvoreanu M, Heger A, Passi S, Dhanorkar S, Kahn Z, Wang R (2023) Responsible AI maturity model. Technical report MSR-TR-2023-26, Microsoft
Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I (2023) Sociotechnical safety evaluation of generative AI systems. arXiv preprint arXiv:2310.11986
Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH (2022) Emergent abilities of large language models,
xAI (2024) Open release of Grok-1
Zhou L, Moreno-Casares PA, Martínez-Plumed F, Burden J, Burnell R, Cheke L, Hernández-Orallo J (2023) Predictable artificial intelligence
Zhu K, Wang J, Zhou J, Wang Z, Chen H, Wang Y, Xie X (2023) PromptBench: towards evaluating the robustness of large language models on adversarial prompts