Regulatory Policies on Ethics Evaluations for Large-Scale AI Systems

Neha R. Gupta

Carnegie Mellon University
nehargupta@cmu.edu

Introduction

Machine learning (ML) model evaluation typically focuses on estimating errors of prediction or estimation via quantifiable metrics. Given the increasing size and complexity of large-scale ML systems, comprehensive evaluations are growing more multifaceted, and evaluation metrics are being expanded beyond performance to include factors like fairness, privacy loss, or other harms induced by the machine learning system. The potential ethical harms of large ML systems, and real instances of algorithmic injustice, have garnered attention from practitioners and observers, leading to a growing chorus of calls for more research and development of community standards around system evaluations and audits (Gupta, Hullman, and Subramonyam 2024).

A large breadth of research has proposed new metrics, measurement techniques of bias and safety, and comprehensive documentation frameworks for practitioners to use (Mitchell et al. 2019; Raji et al. 2020, 2022). These proposals aim to reduce the risk of undesirable outputs from systems, but due to the increasing size and complexity of algorithmic systems, unknown risks and uncertainties in likelihood of risks have become intrinsic. This fundamental problem leads to several places for practitioners to stumble while selecting ethical evaluation practices (Gupta, Hullman, and Subramonyam 2024). Further, while the creation of community standards and governance structures based on voluntarism indicates intrinsic motivation towards ethical AI, some might question why ML practitioners are motivated to perform ethical evaluations, and whether there are limits to accountability based on voluntary participation.

In response, a patchwork of regulatory proposals relying on audits have emerged, which preside over varying jurisdictions, and domains of applicability. There have also been calls for comprehensive approval systems (UN 2024). Ultimately, changes to regulatory policies around AI evaluations that impact organization decision making will also impact models. By exploring the relationship dynamics between regulators and practitioners, as well as the competing priorities involved, we can improve our ability to ensure deployed systems are ethical. My research aims to shape a responsible AI future by helping facilitate discussion among stakeholders on the tradeoffs and competing interests involved in AI regulation.

Background

AI Audits

Efforts to move AI systems towards ethical compliance are grounded in a growing body of literature, that has expanded the set of properties functioning systems should possess beyond accuracy metrics, to include values such as non-maleficence, fairness, societal impact, and sustainability (Jobin, Ienca, and Vayena 2019). Fueled by several high-profile instances of algorithmic injustice, a body of research that develops principals and guidelines for testing of AI systems across these values has emerged.

An evaluation of an ML system is a process in which practitioners detect differences between desired and actual model behavior. Testing of systems can take a range of forms, and requires significant decision making on the part of ML practitioners, such as when to make evaluations during the product development lifecycle, and how to report results (Zhang et al. 2020). Comprehensive system evaluations include audits, which are tools to interrogate complex processes, and are often used to determine whether the system complies with industry standards or regulations. Artifacts that a comprehensive audit could yield include "ethical risk analysis charts", which would consider the likelihood of a failure and potential severity (Raji et al. 2020, 2022). There are inevitably challenges in advancing accountability, or the state of being responsible for a system and its impacts, due to difficulties in estimating these likelihoods.

Regulation

The concern that the existence of a well executed audit process alone is insufficient to ensure accountability, has shifted attention to regulatory checks and balances. Recent regulations that incorporate evaluations include, for example, the EU AI Act, which carries audit requirements (European Commission 2024), and the Biden Executive Order on AI, which required that teams developing AI systems engage in red-teaming their systems prior to its rescission (Biden 2023; O'Brien 2025). Other regulatory proposals include discussions around the possibility of the US creating an FDA-like AI approval board (Carpenter and Ezell 2024).

The creation of AI safety policies is challenging, due to a range of issues regulations need to account for, such as the gap between pre-deployment evaluations and failures discovered post-deployment. This prompts a need for regulatory frameworks that are flexible and future-proof (Bengio et al. 2025).

Research Approach

The fundamental question motivating my research is: *How can practitioners and regulators move the ML industry towards accountability and prevention of ethical harms, despite limitations to system evaluations?* Despite carefully detailed work on audit frameworks and industry standards, we ultimately need to move towards accountability mechanisms that account for intrinsic uncertainty and conflicting interests. Thus, *my research focuses on investigating potential implications of industry standards or regulatory mandated audits, examining the factors that shape these regulations, and providing a framework to guide community organization and policy proposals.*

Studying Harms from Audits Themselves

Even when evaluation metrics include factors like fairness, privacy loss, or other harms induced by the machine learning system, this is often focused on the ethical harms of the released system, overlooking possible harms incurred during the machine learning development lifecycle itself. This is problematic because evaluation approaches have the potential to cause ethical harm during evaluation. In a noteworthy example, Tesla's autonomous vehicle live testing systems on public roads has been widely criticized for being involved in various crashes (Kolodny 2021).

My previous research provided a conceptual framework that cast the primary trade-off in ethical evaluation decision-making as balancing the goal of optimizing for information gained in an evaluation, against the possible ethical harms that are induced. Our sketch of this fundamental problem that practitioners face leaned on economic utility arguments. With this utility model, we illustrated challenges that can cause practitioners to stumble in selecting ethical evaluation practices, referencing real-world examples of machine learning evaluations, and drawing parallels between these challenges and evaluation practices in domains other than machine learning, to explore potential mitigation techniques (Gupta, Hullman, and Subramonyam 2024).

Modeling Regulation

The work in (Gupta, Hullman, and Subramonyam 2024) focused on a team's evaluation decision. If regulations are enforced on audits or evaluations, these would constrain team's utility maximization in decision making.

Since governing bodies creating regulations and ML practitioners have different stakeholders interests in mind and different utility functions, I have been exploring the study of enforcement strategies and stochastic externalities from contract theory and economics, which might cast this problem as a principal-agent model (Bolton and Dewatripont 2004; Cohen 1987). This could use a two-mechanism approach to corporate compliance (in which one mechanism confirms regular evaluations are performed, and a second penalizes instances of harm), thus accounting for multiple concerns held by the ML safety community: The need for monitoring to identify risks, and the need to evaluate performance once general purpose AI systems are already in use (Bengio et al. 2025). I also explore how policy discourse may lend insights into the economic modeling of this problem.

Ultimately, the motivation to designing such a framework is to allow exploring whether an optimal enforcement contract can be created. If regulators set terms via penalty amounts, it is possible they can ensure the penalties are selected such that the ML teams choose the regulator's preferred evaluation efforts.

Exploring Statistical Significance in Audits

Audit processes, and regulatory mandates, fall short in their goal of guaranteeing fairness of machine learning systems, due to the poor statistical rigor mandated in audits. In one example, the New York Local Law 144 (LL 144), which mandates third party bias audits on automated resume screening software used in New York, demonstrating compliance involves offering summary statistics of the percentage of candidates selected by their screening algorithm from each race and gender category (New York City Department of Consumer and Worker Protections 2023). It could be argued that this is an insufficiently rigorous causal claim of an algorithm's lack of biases; There is no indication that additional features of candidates that might create systemic differences, such as education levels, have been controlled for. The LL 144 has been criticized for various loopholes that allow evasion of mandatory audits due to the broad definition of a resume screening system (Groves et al. 2024), and in addition to this, I hope to address the reporting style of audit results.

Some statistically convincing audit procedures have been recommended in prior research. For example, a fair dummies test has been proposed, in which models provide a p-value for satisfying the equalized odds property (Romano, Bates, and Candes 2020). The use of modified conformal prediction intervals could also move towards better reporting of predictions and easier audits (Romano et al. 2020). In my future work, I hope to explore these in more depth, and explore using the reliability of the statistical interpretation of an audit result as a parameter to a regulatory decision process. I also have an interest in investigating issues in resume screening regulation in more depth, with interest in how the ultimate policy that emerged was influenced by various stakeholders.

References

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; Heidari, H.; Ho, A.; Kapoor, S.; Khalatbari, L.; Longpre, S.; Manning, S.; Mavroudis, V.; Mazeika, M.; Michael, J.; Newman, J.; Ng, K. Y.; Okolo, C. T.; Raji, D.; Sastry, G.; Seger, E.; Skeadas, T.; South, T.; Strubell, E.; Tramèr, F.; Velasco, L.; Wheeler, N.; Acemoglu, D.; Adekanmbi, O.; Dalrymple, D.; Dietterich, T. G.; Felten,

E. W.; Fung, P.; Gourinchas, P.-O.; Heintz, F.; Hinton, G.; Jennings, N.; Krause, A.; Leavy, S.; Liang, P.; Ludermir, T.; Marda, V.; Margetts, H.; McDermid, J.; Munga, J.; Narayanan, A.; Nelson, A.; Neppel, C.; Oh, A.; Ramchurn, G.; Russell, S.; Schaake, M.; Schölkopf, B.; Song, D.; Soto, A.; Tiedrich, L.; Varoquaux, G.; Yao, A.; Zhang, Y.-Q.; Ajala, O.; Albalawi, F.; Alserkal, M.; Avrin, G.; Busch, C.; de Carvalho, A. C. P. d. L. F.; Fox, B.; Gill, A. S.; Hatip, A. H.; Heikkilä, J.; Johnson, C.; Jolly, G.; Katzir, Z.; Khan, S. M.; Kitano, H.; Krüger, A.; Lee, K. M.; Ligot, D. V.; López Portillo, J. R.; Molchanovskyi, O.; Monti, A.; Mwamanzi, N.; Nemer, M.; Oliver, N.; Pezoa Rivera, R.; Ravindran, B.; Riza, H.; Rugege, C.; Seoighe, C.; Sheehan, J.; Sheikh, H.; Wong, D.; and Zeng, Y. 2025. International AI Safety Report. Technical Report DSIT 2025/001.

Biden, J. R. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.

Bolton, P.; and Dewatripont, M. 2004. *Contract theory*. MIT press.

Carpenter, D.; and Ezell, C. 2024. An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 239–254.

Cohen, M. A. 1987. Optimal enforcement strategy to prevent oil spills: An application of a principal-agent model with moral hazard. *The Journal of Law and Economics*, 30(1): 23–51.

European Commission. 2024. The EU Artificial Intelligence Act. https://artificialintelligenceact.eu/ai-act-explorer/.

Groves, L.; Metcalf, J.; Kennedy, A.; Vecchione, B.; and Strait, A. 2024. Auditing work: Exploring the New York City algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1107–1120.

Gupta, N. R.; Hullman, J.; and Subramonyam, H. 2024. A Conceptual Framework for Ethical Evaluation of Machine Learning Systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 534–546.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Kolodny, L. 2021. A federal agency warns Tesla tests unfinished driverless tech on its users. https://www.cnbc.com/2021/03/12/tesla-is-using-customers-to-test-av-tech-on-public-roads-ntsb.html. Accessed 2025-08-19.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

New York City Department of Consumer and Worker Protections. 2023. Automated Employment Decision Tools (Updated). https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/. Accessed: 2025-07-19.

O'Brien, M. 2025. Trump rescinds Biden's executive order on AI safety in attempt to diverge from his predecessor. https://apnews.com/article/trump-ai-repeal-biden-executive-order-artificial-intelligence-18cb6e4ffd1ca87151d48c3a0e1ad7c1. Accessed 2025-08-19.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571.

Romano, Y.; Barber, R. F.; Sabatti, C.; and Candès, E. 2020. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2): 4.

Romano, Y.; Bates, S.; and Candes, E. 2020. Achieving equalized odds by resampling sensitive attributes. *Advances in neural information processing systems*, 33: 361–371.

UN. 2024. Governing AI For Humanity: Final Report.

Zhang, J. M.; Harman, M.; Ma, L.; and Liu, Y. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1): 1–36.