Disclosure and Evaluation as Fairness Interventions for General-Purpose AI

Vyoma Raman[12] **Judy Hanwen Shen**[1] **Andy K. Zhang**[1] **Lindsey Gailmard**[1] **Rishi Bommasani**[1]
Daniel E. Ho[*1] **Angelina Wang**[*12]

[1]Stanford University
[2]Cornell Tech

vr359@cornell.edu

Abstract

As generative models are increasingly deployed across diverse settings, fairness interventions must move beyond model-level risk to address system and society-level risk. This expansion is challenging because general-purpose AI (GPAI) is used widely and potential use cases are often not known in advance. As such, we ask: *what fairness-related requirements are feasible when fairness is inherently contextual but we lack context for GPAI*? We specifically consider the obligations of two major groups: system providers and system deployers. While system providers are natural candidates for regulatory attention, the current state of AI understanding offers limited insight into how upstream fairness harms translate into downstream impacts. Since usage contexts are unknown and fairness is not universally defined across them, rather than imposing context-agnostic requirements, we instead argue for transparency from providers such as to whom they are serving their models and a better understanding of how model development decisions influence fairness. On the other hand, system deployers are closer to real-world contexts. Even if they still lack comprehensive knowledge of all use cases, they can leverage their proximity to end users to address fairness harms in different ways. Here, we argue they should responsibly share information about users and personalization and conduct rigorous evaluations across different levels of fairness. Overall, instead of focusing on enforcing outputs from each group (e.g., that a model has selection rates that are at least 80% representation of each other), we propose a *process*-oriented breakdown of obligations between system providers and deployers centered on systematic data collection. This allows us to be specific and concrete about the processes even while the contexts remain unknown. Ultimately, this approach can sharpen how we distribute fairness responsibilities and inform more fluid, context-sensitive interventions as AI continues to advance.

Introduction

As AI is applied to new domains and deployed in new contexts, the potential for fairness-related harms grows significantly across multiple levels. These harms do not remain isolated; rather, biases can cascade and amplify inequities across each level, underscoring the complexity and urgency of addressing fairness in AI systems.

Research illustrates how fairness harms can emerge and escalate from various points across three levels of AI deployment: the model level, system level, and society level (Suresh and Guttag 2021). At the model level, biases in training data or algorithmic design can produce disparities across demographic groups. For example, melanoma detection models often demonstrate higher accuracy for lighter skin tones (Daneshjou et al. 2022; Montoya, Roberts, and Hidalgo 2025). At the system level, these biases can intensify when the model is integrated into decision support systems. When AI is deployed in domains like policing or hiring, outputs are actively recontextualized by human decision-makers who hold different levels of skepticism and agency that affect potential inequities (Brayne and Christin 2021; Kiviat 2018). At the society level, the compounded effects of biased decision-making tools can exacerbate structural inequities. For instance, in medical diagnosis, persistent inequalities in critical predictive attributes like race or health cost (Obermeyer et al. 2019; Eneanya, Yang, and Reese 2019) can combine with underperforming AI systems. This may result in reduced access to timely and accurate treatment for underserved populations, further entrenching inequities in healthcare delivery. At the system and society level, the context where the AI is being applied shapes its ultimate effects.

A key refrain in fairness literature is that researchers cannot "treat fairness and justice... separate from a social context" (Selbst et al. 2019). This principle informs everything from which metrics to choose to which social groups to consider to how to determine and represent relevant factors of individuals. However, the emergence of general-purpose AI models (GPAI), often referred to as foundation models (Bommasani et al. 2021), complicates this approach. GPAI is characterized by its applicability to a wide range of tasks, many of which may be unforeseen at the time of development. Within the AI lifecycle, we distinguish between two primary roles: system *providers* and system *deployers*. System providers are entities that make GPAI models accessible to others, either by distributing them directly or by exposing them through their own interfaces. This definition includes those who release pretrained models (e.g., Meta with Llama), provide platform access to models via APIs (e.g., OpenAI, Anthropic), or facilitate model sharing through hubs (e.g., Hugging Face, Together AI). What qual-

[*]Equal senior authorship.

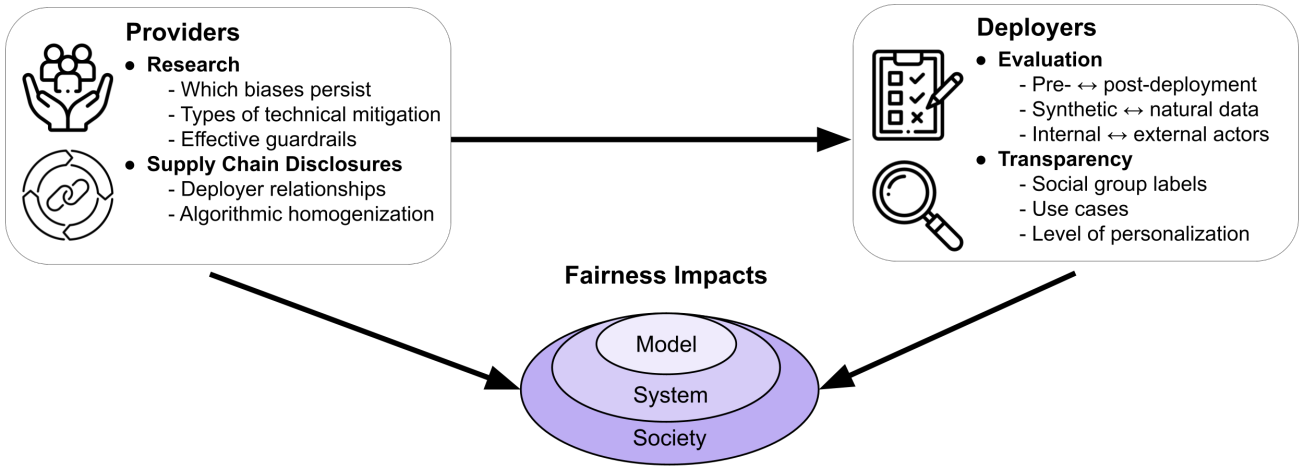

Figure 1: We recommend fairness-related obligations involving information-gathering for providers and deployers to help us analyze and mitigate harms at the model, system, and society level.

ifies an entity to be a provider is not whether they developed or fine-tuned the underlying model, but rather that they made it available for external use in a general-purpose capacity, either in its original or modified form. The obligations we place upon system providers can also be satisfied by those who develop the models. [1] System deployers, by contrast, are entities that integrate a model into end-user-facing or internal applications. This includes incorporating GPAI into customer-facing tools like chatbots (e.g., Klarna's AI assistant using OpenAI's models), embedding models in productivity platforms (e.g., Microsoft's Copilot), or leveraging them internally for tasks such as document summarization, code generation, or predictive analytics. These definitions align with the EU AI Act (European Parliament and Council 2024). We examine fairness-related obligations for AI system providers with limited foresight into post-deployment modifications and uses, and for deployers who may not fully anticipate eventual applications. We delineate the respective responsibilities of the providers and deployers of AI systems and how they address different levels of impact, as shown in Figure 1.

Despite growing attention to fairness in generative AI, existing frameworks often struggle to assess and mitigate harms across diverse and shifting contexts—particularly when systems are designed for general-purpose use and deployed in unforeseen ways. Much of the current scholarship on generative AI focuses on representational harms, examining how groups or individuals are misrepresented in AI outputs (Barocas et al. 2017; Katzman et al. 2023; Intelligence 2024; Teo, Abdollahzadeh, and Cheung 2023; Ghosh, Lutz, and Caliskan 2024; Hofmann et al. 2023). While these studies include sophisticated analyses of subtle and structural biases, they often emphasize model-level attributes (e.g., linguistic or visual patterns), which may not fully capture how these representations function in real-world deployments or interact with broader sociotechnical systems. This gap becomes more pronounced in GPAI, where the lack of predefined use cases makes it harder to anticipate how harms might emerge across different applications (Ferrara 2024; Gallegos et al. 2024; Weidinger et al. 2022; Cohen et al. 2025). Depending on the context, harms may extend beyond biased representations to include issues such as information leakage, privacy violations, and the spread of disinformation or toxic content (Weidinger et al. 2022; Xiang 2024; Kearns 2024). These cascading harms demand fairness frameworks sensitive to both representational concerns and deployment contexts.

Given the seeming paradox of fairness requiring contextual understanding, and general-purpose AI lacking context, we emphasize forms of systematic information-gathering to support more subsequent context-sensitive responses. Specifically, information is needed about users, supply chain relationships, and output distributions. Currently, many fairness efforts focus on visible or well-documented issues such as biased representations or disparities in performance metrics. While important, it is unclear how to apply this information to effectively mitigate fairness issues at different levels. Information-gathering can itself act as a fairness intervention by increasing transparency, enabling accountability, and creating market pressure for companies to adopt more ethical practices (Wang, Datta, and Dickerson 2024). More importantly, it lays the groundwork for adaptive governance by allowing researchers, policymakers, and civil society actors to better anticipate how fairness harms emerge, accumulate, or shift over time. By collecting and analyzing data from across the AI development and deployment pipeline, we can develop more targeted, flexible interventions that are responsive to the dynamic nature of GPAI systems.

Our Work

In this work, we propose an approach to addressing fairness-related harms in GPAI, emphasizing the role of gathering information on GPAI and its deployment contexts as a critical

[1] It may be worth taking different approaches to defining this group as new research grows our understanding of how developer decisions (e.g., training data curation) affect downstream tasks.

intervention strategy in contexts where future use cases and model capabilities are uncertain. In the following sections, we:

- **Refocusing Fairness Interventions:** We draw on prior work that has broadened the scope of fairness beyond the model level to focus on the system and society level. We argue for increased information-gathering to better understand the scope of risk at these higher levels, as well as how to intervene.
- **Provider Obligations:** We articulate fairness-related responsibilities for system providers, including supporting research ecosystems and increasing transparency around model recipients and provenance. In doing so, we challenge the common intuition that fairness can be ensured through dataset curation or universal disparity mitigation.
- **Deployer Obligations:** We outline deployer obligations under broad deployment conditions, emphasizing increased transparency and diverse evaluations that can better characterize use.
- **Operationalizing Fairness for Regulatory Contexts:** We offer recommendations on setting guidelines for which models and systems are subject to particular types of regulation, and we discuss alignment with ongoing regulatory and political developments.

Refocusing Fairness Interventions

We contextualize information-gathering as a critical mechanism for surfacing fairness harms across the model, system, and society levels of GPAI systems. First, we describe how information-gathering can operate across three levels of fairness. Next, we examine how current legislation addresses and distributes responsibility for fairness in GPAI across actors and argue for a reallocation of obligations. Finally, we describe the role of information-gathering in enhancing fairness interventions.

Levels of Fairness

As AI systems diffuse across sectors and are adapted to increasingly diverse applications, the relative severity of fairness-related harms shifts from model-level disparities to broader system and society impacts. While prior work has emphasized the need to broaden fairness analyses beyond predictive disparities (Green and Hu 2018; Selbst et al. 2019; Wang et al. 2022), these broader harms have often been overlooked or under-measured, especially as GPAI is repurposed for unanticipated uses where existing benchmarks based on model outputs may inadequately measure risks. Addressing these cascading harms requires a multi-level approach that situates fairness not only within model outputs but also within institutional workflows and socio-economic structures.

Model Level. Model-level fairness analyses quantify disparities in outputs and predictions that result from datasets, model architectures, and training practices. These studies focus particularly on representational and allocational harms (Barocas et al. 2017), such as racial disparities in medical diagnostic models or gender biases in hiring algorithms. For instance, models for melanoma detection are more accurate for lighter skin tones due to the overrepresentation of lighter-skinned patients in training datasets (Daneshjou et al. 2022; Montoya, Roberts, and Hidalgo 2025). Similarly, automated speech recognition (ASR) systems consistently underperform for speakers of African American Vernacular English (AAVE) (Koenecke et al. 2020). While model-level analyses can effectively surface such disparities, they often focus on immediate, quantifiable harms. The evaluations provide only a preliminary look at biases, and harms may manifest differently or propagate more broadly in real-world contexts. Additionally, such analyses typically assume clearly defined use cases. This assumption is increasingly untenable in the context of GPAI, where models are deployed across multiple, unforeseen applications.

System Level. System-level fairness focuses on how AI outputs interact with human decision-makers and organizational processes, particularly where individual-level decisions are made within structured organizations. Unlike model-level analysis, which centers on output disparities, system-level analysis addresses how those outputs shape specific decisions—such as in hiring, lending, or legal judgments—within real-world workflows. For example, ASR models that exhibit linguistic biases may impact decisions that are made based on resulting transcripts, such as when applied to courtroom transcription (Prasad et al. 2002; Martin and Wright 2023). If transcripts with systematic errors for particular social groups become part of official records, these inaccuracies can compound existing legal disparities, affecting judicial outcomes and access to justice. Similarly, predictive models used for hiring, lending, or law enforcement may inadvertently reproduce or even amplify existing biases within organizational systems (Cohen et al. 2025; LangChain 2025). Identifying system-level harms presents greater methodological complexity than model-level analyses, as these impacts are dependent on both context and human operators of technology. Effective analyses at this level require information-gathering that captures interactions between model outputs and institutional practices, highlighting the need for ongoing, context-sensitive evaluations.

Society Level. Society-level fairness examines how the cumulative effects of decisions made with AI in the system reshape broader patterns of inequality across populations. Rather than focusing on individual harms, this level particularly assesses how AI integration influences the long-term distribution of resources, opportunities, and risks. For example, in healthcare, models that underdiagnose disease by applying historical proxies like medical cost data and race-adjusted metrics (Seyyed-Kalantari et al. 2021; Obermeyer et al. 2019; Eneanya, Yang, and Reese 2019) can compound to lower the quality of care received by marginalized populations overall. In education, AI-based learning tools may be applied to cover gaps in underfunded school systems (Relan 2025; Sylvestre 2025) despite questionable impact on learning rates (Bastani et al. 2024), resulting in inconsistent quality of education. In labor markets, AI-driven automation and decision-making systems can displace workers or

shift them into more precarious positions, intensifying wage suppression and economic insecurity (Capraro et al. 2024). While all of these issues can individually occur at the system level, society-level fairness asks whether the existence of AI in a particular ecosystem reifies disparities in health, wealth, and power. This broader framing requires an analytical approach that considers how AI systems degrade or enhance resources, who retains access to higher-quality services, and how these dynamics shape broader patterns of inequality over time. This can involve a variety of work, including economic analyses like the Anthropic Economic Index (Handa et al. 2025b) that tracks the distribution of benefits and harms as AI becomes integrated into critical sectors, or efforts to evaluate impacts of AI and fairness interventions broader societal wellbeing over time (e.g., Liu et al. 2018). Beyond sector-related trends, society-level fairness also entails examining how the energy consumption and carbon emissions of AI systems are distributed (Luccioni et al. 2024). This may disproportionately affect communities with fewer resources and less resilience to climate impacts. Such assessments are only possible if providers and deployers are transparent about use cases, deployment contexts, and system integration. Accordingly, we emphasize the need for actors to contribute to information-gathering across model, system, and society levels to comprehensively assess the distributional consequences of GPAI systems.

Actors and Priorities in AI Regulation

We delineate the distinct responsibilities of system providers and system deployers. While the AI lifecycle envisions the roles at different stages, entities may assume dual responsibilities depending on their level of control and intervention in the model's development and deployment. This categorization of actors is particularly relevant in light of regulatory frameworks like the EU Artificial Intelligence Act, which categorizes providers as entities who develop or place AI systems on the market and deployers as those utilizing AI systems under their authority beyond personal, non-professional use (European Parliament and Council 2024). Importantly, the boundary between these roles can blur when a deployer undertakes significant modifications to a GPAI. Such modifications may assign an entity both deployer and provider status under the AI Act, expanding their obligations. By adopting the same approach, we underscore the necessity of holding such entities accountable for fairness-related risks, regardless of their initial role in the AI supply chain. This approach ensures a more comprehensive accountability framework, aligning responsibilities with the actual level of influence an entity exerts over a model's outputs and deployment.

Emerging AI regulatory frameworks have lacked specificity on fairness concerns compared to safety, security, and catastrophic risks. Most approaches apply a similar two-actor framework highlighting multiple intervention points, yet existing policies tend to specifically target acute risks visible at the model or system level rather than pervasive society-level harms (Bernardi et al. 2025). For instance, while the EU AI Act does include non-discrimination provisions, its criteria for high-risk systems emphasize health and safety concerns rather than directly addressing biases and inequities (European Parliament and Council 2024). Similarly, the General-Purpose AI Code of Practice prioritizes mitigating "serious incidents and malfunctions" rather than broader, aggregate impacts of biased AI outputs (European Commission 2025). In the U.S., regulatory efforts such as President Trump's Executive Order 13859 emphasized national security, economic competitiveness, and human flourishing without addressing discrimination or bias explicitly (Executive Office of the President 2025), while other orders like Executive Order 13960 (Executive Office of the President 2020) and President Biden's Executive Order 14110 (now revoked, Executive Office of the President 2023) offered only broad, aspirational references to fairness and civil rights, lacking specific guidance or requirements.

This regulatory gap is further compounded by fairness frameworks that remain largely abstract and decontextualized, emphasizing model-level metrics while overlooking the complex socio-technical dynamics that shape how AI systems interact with institutional processes and user populations (Selbst et al. 2019; Green and Hu 2018). Given their proximity to deployment settings, deployers are uniquely positioned to surface these emergent risks through localized assessments and ongoing monitoring, yet existing regulatory frameworks provide little guidance or accountability for such interventions. Thus, we argue that a process-oriented approach focused on information-gathering by both providers and deployers is essential for detecting and mitigating context-dependent harms.

Information-Gathering as a Fairness Intervention

While information-gathering is often framed as a passive activity, it can also serve as a direct intervention in mitigating fairness harms by surfacing and deterring problematic practices in real time. When system providers and deployers are required to disclose demographic breakdowns of datasets, algorithmic design decisions, or model recipients, they are compelled to confront potential biases and inequities that might otherwise remain obscured. While transparency can prompt such reflections, it is not a guarantee of accountability; without sufficient context and mechanisms for redress, simply revealing information may not prevent harmful practices (Ananny and Crawford 2016). Nonetheless, in a handful of examples, companies have altered practices to avoid reputational damage, regulatory scrutiny, or potential litigation. For instance, algorithmic audits that uncover racial disparities in predictive policing systems have led to public outcry, resulting in the suspension of those systems or substantial modifications to their deployment (Raji et al. 2020). Similarly, well-publicized audits have occasionally led to changes in deployed models (Raji and Buolamwini 2019). By making fairness risks publicly visible, systematic data collection can function as a proactive corrective measure, not just a retrospective assessment.

Our call for information-gathering further seeks to lay the groundwork for future fairness mitigations. Only by better understanding how the different components of GPAI ecosystems interact, such as how the pre-training data of a model can propagate into a downstream fine-tuned model,

can we be more prescriptive to different actors earlier in the pipeline. Our technical understandings of these systems are not yet sufficiently developed, but by increasing research into these interactions, we may be able to eventually allocate additional responsibility to actors earlier in the pipeline.

Provider Obligations

When assigning responsibility for fairness-related concerns in AI systems, it is natural to focus on system providers as key stakeholders. The European Union's AI Act, particularly Article 53, outlines general obligations for providers (European Parliament and Council 2024). In this section, we consider a narrower question: whether providers who are *not* also deployers bear affirmative obligations related specifically to fairness. Our position is grounded in the broader recognition that fairness, like safety, is not an intrinsic property of a model itself but emerges through its deployment and use (Narayanan and Kapoor 2024). We argue that, given current understandings of the bias transfer hypothesis (i.e., that bias in pretrained models will propagate into downstream ones) and dual-use cases, providers' responsibilities are more indirect.

One prior regulatory approach has been to eliminate bias from models and data. Early drafts of the EU AI Act proposed that data used in high-risk systems be "sufficiently relevant, representative and free of errors and complete in view of the intended purpose" (European Parliament and Council 2024). The final version softens this requirement to mandate that datasets be "relevant, sufficiently representative, and to the best extent possible, free of errors and complete," though even determining what constitutes sufficient representativeness is arguably context-dependent (European Parliament and Council 2024). This shift reflects a broader recognition that social biases are relative to the given context and cannot be eliminated when training general-purpose models. For similar reasons, "debiasing" a model is impossible. For example, the dual-use nature of AI systems complicates fairness interventions. A model that has been "de-biased" to avoid producing racially discriminatory language may no longer be usable for socially valuable tasks, such as identifying racially restrictive covenants in property deeds, legal clauses that historically excluded people of certain races from home ownership or occupancy (Surani et al. 2024).

While system providers should address clear and egregious fairness issues—such as datasets where all depictions of a group are inappropriate or offensive—many harms are more subtle. Tools like model cards and dataset documentation are now common mechanisms for provider transparency (Liang et al. 2024). However, we still lack a robust understanding of how training data and modeling decisions shape fairness outcomes in downstream applications. This gap hinders deployers trying to assess risks and undermines the development of enforceable standards for fairness.

We support prior calls for transparency regarding various social considerations in GPAI (Luccioni et al. 2024; Uma Rani 2024; Bommasani et al. 2025), including clarifying environmental harms of model training and deployment and their effect on local communities, ensuring fair labor practices for data workers, and respecting intellectual property rights in the construction of training datasets. While many of these issues are not fairness specific and relate to more general harms, marginalized communities are often disproportionately affected by these system and society-level harms (Gyevnar and Kasirzadeh 2025). This also means that practices that reduce environmental harm, labor exploitation, and power concentration can also mitigate disparities in who is most affected by AI development (Hoes and Gilardi 2025). However, we do not suggest that highly proprietary or implementation-specific information such as fine-tuning recipes or sensitive model internals must be disclosed today. Rather, we argue that systematic, structured information can offer a pragmatic, actionable basis for fairness interventions. For example, reporting regional energy consumption for inference workloads can help identify disproportionate environmental burdens in low-income or climate-vulnerable communities and prompt targeted investment in renewable infrastructure or demand-shifting policies. Over time, as the field matures, deeper insight into the interaction between provider-side decisions and downstream adaptations (including fine-tuning) may support more effective governance. We also highlight two additional areas where systematic information-gathering would be particularly valuable.

Research Investment. To address this, system providers, who control key parts of the development pipeline and generally possess substantial resources, must invest in research that clarifies how their decisions at different stages affect future model outputs. This is essential to develop actionable fairness interventions and build guardrails that remain effective even as models evolve. In particular, providers should prioritize understanding which biases persist through pretraining and fine-tuning pipelines and how they manifest in model outputs (e.g. Kumar et al. 2025). They should also examine how fine-tuning can be used to mitigate harmful biases, considering both the quantity and type of tuning to concretely improve models (e.g., Qi et al. 2025). The research should also analyze and improve the robustness of model-level fairness guardrails (e.g., Qi et al. 2024; Wang and Russakovsky 2023), including through adversarial attempts to produce extremely problematic behavior (e.g., Wallace et al. 2025). Investing in this research not only enables more targeted and effective fairness interventions, but also contributes to the long-term stability of model behavior, fosters broader user adoption, and offers providers a competitive edge.

Supply Chain Disclosures. We argue that an important obligation for providers of GPAI is the disclosure of supply chain relationships. Without visibility into which deployers are using a model and in what domains, it becomes challenging to trace where and how fairness harms emerge and could be addressed. When downstream systems exhibit discriminatory behavior, it is often unclear whether the cause lies in upstream model characteristics, deployment-specific modifications, or contextual misuse. Supply chain transparency provides the necessary infrastructure to attribute responsibility appropriately, enabling researchers, regulators, and im-

pacted communities to identify the relevant actors, investigate causal pathways, and design targeted interventions supported by research. In its absence, providers and deployers may deflect accountability onto one another, stalling harm mitigation. Transparency also supports the conditions under which third-party auditors, civil society organizations, and academic researchers can conduct meaningful evaluations, particularly in high-impact domains where fairness harms may otherwise remain obscured.

Beyond attribution and accountability, supply chain transparency also plays a critical role in enabling fairness at scale. When the same base model is reused across a wide array of applications, its embedded biases can replicate and compound across sectors, resulting in algorithmic homogenization (Creel and Hellman 2022; Bommasani et al. 2022). Individuals may face repeated disadvantages across employment, education, and housing decisions if the same flawed inference patterns follow them from system to system. Disclosing deployment relationships allows auditors to monitor the cumulative effects of model reuse and identifying systemic risks that might be invisible in isolated evaluations. Moreover, such visibility enables anticipatory governance: with sufficient information about where and how models are used, it becomes possible to flag high-risk applications, monitor sensitive domains for overconcentration, and align fairness interventions with actual deployment contexts. While some disclosure obligations may need to be scoped to protect proprietary information, carefully designed transparency regimes—whether through regulators, certification bodies, or consented data sharing—can help balance commercial concerns with the public interest in equitable AI deployment.

Deployer Obligations

While providers may release context-agnostic metrics for model performance, the fairness of general-purpose models depends on how they are used in specific contexts. The EU's General Purpose AI Code of Practice describes the commitments of those involved in the deployment of GPAI systems (European Commission 2025). Specifically, it highlights obligations to formally document and provide relevant information about models and to assess and mitigate potential harms in AI systems, including issues like illegal discrimination and bias in high-risk application areas.

Because the applications of general-purpose models vary, fairness cannot be guaranteed by focusing on specific social outcomes. A model that appears fair in one use case and level may have harmful effects in another, depending on factors such as the population it affects, the decision-making context, and the way it is integrated into broader systems. As a result, fairness must be embedded into AI deployment through procedures that allow for iterative and context-sensitive measurement and development. System documentation and evaluation thus become critical information-gathering approaches to uphold fairness and engage with the ways a model interacts with particular environments.

Transparency

We call for three kinds of transparency that we believe are most important for ensuring AI fairness. Specifically, information about the social groups that users of AI systems belong to, how the system stores and personalizes responses, how users are interacting with the system, and relationships to other actors in the supply chain is critical to evaluating and subsequently improving the fairness of AI systems.

Social Group Labels of Users. First, we echo calls for more transparency and collection of group attribute labels (Bogen, Rieke, and Ahmed 2020; Ho 2020). These are needed to determine where a model underperforms so that targeted intervention can help improve overall accuracy. For instance, during machine learning training, it is common practice to search for the "hardest" training labels. Collecting group attribute labels can serve as a proxy to help identify subsets of the data that a model is not performing as well on. However, it is not always safe for members of marginalized communities—such as non-citizens and queer communities—to share their identities, and inferring or collecting the data may place the communities at increased risk of surveillance (Tomasev et al. 2021; Wachter 2020; Bogen 2024). To safeguard privacy, organizations should aim to follow siloed processes to collect, access, and analyze the data and to minimally determine the attributes of interest (Recommendation: National Artificial Intelligence Advisory Committee (NAIAC) 2024; King et al. 2023).

Personalization. Next, we call for transparency about the type of personalization employed by GPAI. It is important to understand how much of a user's interaction history is recorded and used. For example, DeepSeek reportedly collects user interaction data on servers in China, and Snapchat uses user chat history to personalize recommendations (Newman and Burgess 2025; Snap Inc. 2023). Released human-LLM datasets have been found to contain personally identifiable information (PII), detailed sexual preferences, and specific drug use habits (Mireshghallah et al. 2024), which users may not have intended to make public or accessible for personalization purposes. These privacy risks can also include phone numbers and home addresses of individuals. Transparency in how personalized advertising is generated with user data has remained limited due to the complexity of cookie-based tracking across multiple online platforms. It is essential to provide clear and accessible information about whether companies retain interaction history, and if so, how it is utilized.

Transparency in how personalization works is critical not only for privacy protection, but also for distinguishing personalization and stereotyping. For instance, there are documented concerns about stereotyping based on characteristics such as a person's name (Wilson and Caliskan 2024). The release of system prompts in particular can provide insight into how chatbots are instructed to use different types of information and avoid problematic content. For instance, DeepSeek has released their system prompt saying "Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature"

(DeepSeek-AI et al. 2024). Increasing public knowledge of the internal workings of GPAI arguably reduces security risks (Hall, Mundahl, and Park 2025). Similar to GDPR, transparency around personalization should also be dual-layered, containing the type of personalization made public and the specifics accessible to each user.

Use Cases of AI Systems. Finally, we echo previous calls for transparency around the use cases and supply chain of AI models (e.g., Zhao et al. 2024). Without information about how AI systems are used in practice, it is hard to understand how to evaluate them and which actors are best positioned to do so. Contexts like summarizing patient records or answering students' questions about academic material (e.g., Handa et al. 2025a) can inform the specific tests used to evaluate the model as well as the tolerance level for failures. However, usage-based evaluations are only possible with transparency of use cases. This includes not only knowing how a model is used but also whether it has been modified from its original form, as such changes may significantly affect fairness outcomes in deployment. In 2020, U.S. Executive Order 13960 established the Agency Inventory of AI Use Cases (Executive Office of the President 2020). While this policy underscores the value of cataloging AI use, expanding such disclosure to private companies poses practical and legal challenges, particularly given the difficulties already encountered by federal agencies. Still, targeted transparency about both use cases and model modifications—such as in high-risk domains or through voluntary standards (e.g., Editors of HLR 2024)—could help support downstream fairness evaluations without imposing excessive burdens.

Overall, from deployers we advocate for three forms of transparency—group labels, personalization levels, and use cases—that can help users make informed decisions about which models to use based on how they may be treated as a result. Group and use-case information particularly illuminates AI's economic impact across occupations and demographics while creating market incentives for responsible governance. These disclosures need not be fully public: demographic data can be reserved for auditors and agencies, and use-case or supply chain information, while potentially sensitive, carries limited consumer risk and may even benefit firms reputationally (Kraft and Zheng 2021). Still, market incentives alone are insufficient; regulation is needed to ensure companies disclose socially valuable information to relevant stakeholders. Although some business relationships are already subject to disclosure (e.g., SEC filings), they remain difficult to access. Together, these forms of transparency can improve evaluation by enabling auditors to better approximate real-world deployment conditions.

Evaluation

With adequate transparency about the usage and development of AI systems, it is possible to conduct robust evaluations. These, too, are an information-gathering mechanism that identifies where and how a model produces distinct output distributions, enabling targeted unfairness mitigation strategies and informing responsible deployment. We identify three important design decisions that should be considered when mandating evaluations or scoping models based on the results of evaluation. These include whether the evaluation is done: (a) before and after model deployment, (b) data that reflects natural and unnatural (i.e., synthetic) contexts, and (c) from within and outside the developer organization. While both sides of each of these axes are valuable, they are not implemented to the same degree and have different implications for the levels of fairness. Thus, we recommend a balanced approach to bring fairness back to the fore of evaluation.

Evaluation	Stage	Data	Actor
Benchmarks	Pre-deploy	Synthetic	Internal
Bug bounty	Pre-deploy	Synthetic	External
Historical data	Pre-deploy	Natural	Internal
Compliance audit	Pre-deploy	Natural	External
Simulations	Post-deploy	Synthetic	Internal
Public red teaming	Post-deploy	Synthetic	External
AB testing	Post-deploy	Natural	Internal
Incident reporting	Post-deploy	Natural	External

Table 1: Evaluation categorization by deployment stage, data type, and actor performing the audit. No single evaluation covers all fairness concerns; rather, this typology shows that complementary evaluations are needed to identify and mitigate diverse harms.

Pre-deployment vs. Post-deployment. We distinguish between different temporal stages of evaluation, pre-deployment and post-deployment, not as a binary, but as a continuum of iterative oversight. In pre-deployment, an AI system is evaluated before being released to users. This evaluation often prioritizes internal performance standards, typically focusing on technical benchmarks, risk analysis, and compliance with field norms. One common form of pre-deployment evaluation is benchmarking, where suites such as HELM (Liang et al. 2023) assess models along multiple dimensions including accuracy, robustness, calibration, and fairness. HELM aims to provide a more comprehensive evaluation than traditional leaderboards by including a broad range of scenarios and tasks designed to capture different aspects of model performance, including some fairness-related concerns. Another example is red teaming, in which adversarial inputs are designed to elicit harmful outputs from models (Perez et al. 2022). Red teaming efforts often focus on areas like cybersecurity, biosecurity, and content safety, but are generally scoped to harms anticipated by system providers prior to release.

However, pre-deployment evaluations are limited by their reliance on assumptions about user behavior and model usage. Benchmarks are constrained to predefined tasks and may not capture fairness-related harms that arise in diverse real-world settings. Even Dynabench, which introduces dynamic data collection by allowing models to be tested and improved through adversarial user inputs over time, is still limited because it relies on the specific tasks and populations involved in the benchmark construction (Kiela et al. 2021). Projects driven by specialized online communities have his-

torically experienced challenges in recruiting marginalized individuals, e.g. Wikipedia editors (Hill and Shaw 2013). As a result, it may fail to identify fairness failures that arise in new contexts, affect groups who were not represented among early users or adversaries, or manifest only after prolonged real-world interaction (e.g., Associated Press 2025)[2]. Pre-deployment efforts provide important baselines but are insufficient to ensure ongoing fairness once GPAI is widely used. Post-deployment evaluation, on the other hand, assesses GPAI based on how users actually interact with it in the real world. This allows evaluation to address discrepancies between intended and actual use, such as differences in the types of questions asked of general-purpose models and reasoning models. Post-deployment evaluation systems can capture emergent user and model behaviors, novel use cases, and biases that would not have been surfaced during pre-deployment testing.

Different approaches to post-deployment evaluation exist. One is adverse event reporting (AER) systems, which allow users to flag harmful or problematic outputs encountered in practice (Committee 2023). While AER is crucial for identifying significant incidents, it is both reactive and inherently limited by users' ability to recognize and report harm in the moment. Individuals impacted by pervasive harms may not immediately realize that they have been affected. Incident aggregation efforts such as the AI Incident Database (2025), which collects and catalogs instances of AI failure, can identify a subset of broader patterns over time, even as they are constrained in scope by the nature of problems that get reported. Such databases offer another layer of analysis by systematically reviewing the totality of user interactions to detect fairness failures at scale (Dai et al. 2025).

The distinction between pre-deployment and post-deployment evaluation is critical from a fairness perspective because each stage surfaces different types of harm across the model, system, and society levels. Pre-deployment evaluations like benchmarking primarily target model-level disparities, identifying performance gaps in controlled conditions. However, post-deployment evaluations like incident reporting are essential for detecting system-level failures, such as misalignments between outputs and institutional workflows, and society-level harms that emerge over time, like unequal access to services or labor displacement. These broader impacts often affect marginalized groups in ways that cannot be fully anticipated during development. A fairness-centered evaluation pipeline must therefore include both rigorous pre-deployment testing and sustained post-deployment monitoring to capture evolving, context-dependent harms.

Naturalistic vs. Non-naturalistic Assessment. We define naturalistic assessments as benchmarks that draw from real-world data sources and reflect authentic, ecologically valid scenarios (De Vries, Bahdanau, and Manning 2020). Such assessments are especially critical for system and society-level fairness harm detection because they directly measure the risks, harms, and disparities that models may propagate in practice (Shen and Guestrin 2025). In contrast, non-naturalistic assessments that rely on synthetic data generation or controlled perturbations are generally limited to detecting model-level harms, such as output disparities across demographic attributes. While these evaluations provide important insights into model behavior under controlled manipulations, they often lack the contextual richness necessary to capture systematic and structural biases. Prior work has shown that simple perturbations, such as switching demographic attributes in text, can produce absurd or misleading results if real-world social dynamics are not considered. Blodgett et al. (2021) famously illustrate this with the "Norwegian salmon" example, where perturbation-based measures of stereotyping confound nationality and race, leading to spurious conclusions. Naturalistic evaluation guards against such failures by grounding fairness measures in the actual ways that identity, language, and power interact in real environments.

Beyond sociolinguistic tasks, the importance of ecological validity has been highlighted in high-stakes domains like cybersecurity. For instance, the BountyBench framework evaluates language models on three common security tasks: detecting, exploiting, and patching security vulnerabilities. It uses bug bounties rather than synthetic tasks constructed by researchers (Zhang et al. 2025). The findings demonstrate that realistic, complex tasks expose significant model limitations and vulnerabilities, which synthetic researcher-constructed datasets might not. The researchers also attach monetary values to tasks to approximate economic impacts. Thus, authentic benchmarks better reflect the multifaceted, emergent risks and capabilities of AI systems, and metrics should be grounded not just on model outputs but also in approximations of impact and human-level baselines, using historical human data or controlled trials to contextualize what fair and reliable performance should look like.

That said, non-naturalistic data remains essential for evaluating fairness failures at the model level. Controlled perturbations and synthetic counterfactuals allow for fine-grained analysis of specific model behaviors that might be difficult to isolate in naturally occurring data. For instance, counterfactually augmented datasets help models learn to distinguish spurious correlations from causal signals (Kaushik, Hovy, and Lipton 2020). Although such synthetic interventions cannot substitute for real-world grounding, they provide critical tools for stress-testing models and diagnosing fairness failures at a granular level. In a comprehensive fairness evaluation framework, naturalistic assessments should anchor primary evaluations of harm and risk, while non-naturalistic methods should serve as supplementary diagnostics for bias and robustness.

Internal vs. External Evaluation. It is critical to consider the position and incentives of those conducting evaluations when assessing fairness in AI systems. Internal evaluation plays an important role, particularly due to the greater access internal auditors have to models, data, and development teams. Internal teams can conduct more detailed audits, identify model-level fairness failures and system-level repercussions early, and work closely with providers to prioritize fixes (Raji et al. 2020). For example, OpenAI con-

[2]Content warning: Discussion of suicide.

ducted an internal audit examining biases in chatbot interactions by analyzing responses to users of different backgrounds. This study would only be possible through access to sensitive internal data (Eloundou et al. 2024). Anthropic has also analyzed economic implications and values (Handa et al. 2025b; Huang et al. 2025), which provided insight into fairness issues at the interaction level and informed mitigation strategies. However, internal evaluations remain constrained by organizational priorities, incentives, and perspectives, and thus cannot substitute for independent scrutiny.

External evaluators are essential for uncovering issues that might otherwise be deprioritized or overlooked by model development organizations. These differences can occur because external evaluators may lack company-internal assumptions about the users of the AI and what they will do with it. In particular, external evaluation helps surface fairness concerns that may conflict with profit incentives or internal narratives (Longpre et al. 2025). Third-party auditing, where independent actors assess models for biases and misalignments with societal values, serves as a key mechanism for achieving public accountability. However, external evaluators often face limited access to proprietary data and model internals, making it difficult to conduct comprehensive assessments. We echo calls to support external evaluation ecosystems, including providing broader access and legal protections to auditors, to enable evaluations that capture the levels of fairness more expansively (Raji et al. 2022). Such support is necessary to ensure that diverse stakeholders, particularly those most impacted, have channels for oversight.

Building a robust third-party audit ecosystem is central to enabling effective external evaluations. Raji et al. (2022) highlight several critical elements: providing auditors with adequate access to system artifacts, establishing standardized auditing frameworks, and creating safe harbor protections for auditors to mitigate legal risks. Costanza-Chock, Raji, and Buolamwini (2022) similarly recommend resourcing external auditors and mandating public disclosures of audit results, arguing that without independent scrutiny, algorithmic harms—particularly system and societal impacts affecting marginalized communities—are less likely to be surfaced. External evaluation, when properly supported, ensures that fairness concerns at the system and society levels are evaluated from outside the narrow lens of the organizations deploying these systems.

From Evaluation to Impact We raise these three dimensions of evaluation because they highlight important differences between what information is produced and what conclusions can be drawn about fairness at different levels. We do not bring them up to say that one value for each dimension is more important or valuable than the other. For instance, while post-deployment evaluation effectively captures risks that arise in practice, pre-deployment evaluation is critical to ensure that any deployed application has been tested before it is used on real people. At the same time, while naturalistic data is an ideal, using it comes with privacy concerns (Mireshghallah et al. 2024) and issues of scale that synthetic data can help to alleviate. Finally, without the right incentive structures for external evaluators, internal evaluators may currently have the most motivation and access to properly evaluate their models. When creating regulations for evaluation, we need to keep in mind that not all evaluation is equal. Especially for fairness in GPAI where so many of the use cases are unforeseen, we should try to cover the space of possible evaluations.

At the same time, even evaluations that look the same according to our distinctions can differ in critical ways. One evaluation method that is post-deployment, naturalistic, and external is AER, in which users can report incidents of harmful behavior. However, AER fails to capture more subtle, pervasive fairness issues like erasure (Katzman et al. 2023) that cause harm systematically. However, audits based on open-source historical data are also post-deployment, naturalistic, and external but can identify instances of erasure because the data is not pre-filtered for obvious harm.

Ultimately, broader evaluations that span multiple methods and dimensions are not just more informative: they can also drive accountability and change. For instance, when incidents reported through AER require the deployer to implement fixes, this creates a feedback loop that can help deployers and providers to prevent future harm. But more systematic issues, like disparities in how individuals with different names are treated, demand more proactive forms of oversight. These types of harms often go unnoticed unless surfaced through targeted audits or retrospective analyses. Public-facing audits that produce fairness or safety scores could further shift incentives by enabling users to choose services aligned with their values, generating competitive pressure on companies to prioritize equitable behavior. In this way, a more expansive and layered evaluation ecosystem not only reveals where systems fall short but also creates structural levers for improving fairness in GPAI over time.

Operationalizing Fairness for Regulatory Contexts

Effective regulation must be scoped to capture the systems most relevant to its goals while avoiding burdensome overreach on smaller-scale deployers (Laufer, Kleinberg, and Heidari 2025). Poorly scoped fairness regulation risks reinforcing incumbency by making compliance disproportionately easier for large providers. To avoid this, regulators must attend not only to what is regulated, but how scope is defined. We offer two core recommendations for defining what counts as "in scope" for regulation: scope should be determined along multiple dimensions rather than a single threshold, and regulation should apply to systems, not just models. Scoping will necessarily vary depending on whether the regulation targets model developers (where compute might matter more) or deployers (where user scale may be more salient).

Single-threshold rules are insufficient. Compute-based thresholds may serve as a rough proxy for some risks, such as catastrophic misuse, but fail to capture fairness harms, which occur across the full range of model sizes. Even simple models, like logistic regressions used in credit or hiring,

can produce severe discriminatory outcomes (Hooker 2024). In human rights frameworks, which prioritize fairness at the system and society level, risk is categorized based on scale, scope, and remediability (Business for Social Responsibility 2021). Similarly, we disaggregate fairness harms into (1) severity, (2) voluntariness of exposure, (3) scale of deployment, and (4) distribution of harm. For example, organ transplant algorithms may use low-compute models but still warrant intense scrutiny due to high-stakes decisions, lack of user choice, and disproportionate impact on vulnerable groups (Hasjim et al. 2024).

Furthermore, system-level regulation is essential. Many harms emerge not from models alone, but from how they are embedded in decision-making systems involving human discretion. For instance, NYC Local Law 144 exempted systems with human oversight, yet research shows such oversight is often ineffective at preventing bias (Groves et al. 2024; Green 2022). Regulatory focus must therefore extend beyond technical artifacts to include the socio-technical systems in which they are used (Kiviat 2018; Brayne and Christin 2021).

To balance coverage and feasibility, we support calls for dynamic thresholds that evolve with empirical evidence. Overly rigid scoping frameworks, like those seen in the National Environmental Policy Act (NEPA)'s procedural burdens, risk stalling progress (Epstein 2018). Multi-dimensional criteria help avoid overregulating low-impact deployments while ensuring high-impact systems are brought into scope.

Conclusion

In this work, we examine what fairness requires in the context of open-ended generative AI, where the use cases are often undefined at deployment time. While fairness has long been recognized as context-dependent, this raises the question: what obligations are appropriate when the context is unknown? We propose a set of provider and deployer responsibilities grounded in the current technical capabilities and limitations of generative AI. As such, we do not call for disclosures like pretraining data or fine-tuning strategies, both because these are often considered proprietary and because, at present, there is limited methodological clarity on how to meaningfully interpret such information. Future research may establish these disclosures as necessary, but that threshold has not yet been met. Instead, we focus on actionable areas where meaningful oversight is currently possible, such as use case categories, worker conditions, and personalization mechanisms. While maximal disclosure may seem appealing, imposing broad requirements risks burdening smaller actors and further concentrating regulatory power in the hands of large providers and deployers.

When considering GPAI, we are unable to offer the same outcome-based prescriptions that may be done in domain-specific legislation, e.g., specific thresholds on predictive performance disparities that serve as guidance, even if misinterpreted (Watkins and Chen 2024). Instead, we can only prescribe a *process* by which to help regulate fairness-related harms, joining other researchers in prioritizing current and known harms to promote institutional resilience over speculative outcomes (Narayanan and Kapoor 2023). In other words, we prescribe what kinds of transparency and evaluations are important, rather than establishing ex ante evaluation criteria or benchmarks for deployment.

Ultimately, in this work we shed light on the processes rather than outcomes that we should strive for to ensure contextual fairness in purportedly acontextual deployments of GPAI. We make the case that prioritizing regulation, research, and resources to address fairness-related harms is key to leveraging AI for the benefit of everyone.

Adverse Impact

This work offers a focused intervention into fairness-related harms in general-purpose AI, but several limitations and boundary choices are worth acknowledging. First, our regulatory analysis primarily engages with U.S. and European frameworks, such as the EU AI Act and various American executive orders. While these jurisdictions currently shape much of the global conversation around AI governance, our framework does not fully account for legal, cultural, or infrastructural differences in other regions, including Global Majority countries. Future work should examine how fairness responsibilities and data governance norms operate under alternate political conditions and institutional capacities. Second, while we center fairness—particularly in its representational, allocative, and systematic forms—we do not imply that harms like existential risk, misinformation, and ecological sustainability are less important. Our choice reflects a commitment to depth over breadth, with the hope that complementary work will address these parallel challenges.

We also want to emphasize that our call to broaden responsibility beyond system providers is not an attempt to absolve them of accountability for harms stemming from the models they build and distribute. Providers often retain significant control and resources and should be held to strong standards for transparency and impact mitigation. Our argument is that fairness harms frequently emerge from interactions between models and deployment contexts, and that meaningful redress requires shared responsibility. At the same time, we recognize that our proposals—particularly around information-gathering—carry ethical risks of their own. In contexts of weak oversight or coercive governance, increased demands for transparency could be weaponized to deepen surveillance or chill user expression. Similarly, overly burdensome compliance regimes may entrench the power of large incumbents and stifle innovation from smaller actors. Accordingly, any fairness regulation must be paired with robust privacy safeguards, rights-based governance, and attention to how regulatory burdens are distributed. These concerns underscore the need for pluralistic, power-aware implementation and continued reflexivity in how fairness frameworks are operationalized.

Positionality Statement

Our team draws primarily from backgrounds in computer science and law and well-resourced institutions in the United States. As a result, our analysis may prioritize technically feasible interventions that are actionable within current reg-

ulatory and development pipelines. However, we recognize that our perspectives are shaped by our positions within the Global North. As such, we may underemphasize lived experience, localized concerns, or political economies of harm outside U.S. and European regulatory contexts.

Acknowledgements

JHS is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness and the Simons Foundation investigators award 689988.

References

AI Incident Database. 2025. AI Incident Database. https://incidentdatabase.ai.

Ananny, M.; and Crawford, K. 2016. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3): 973–989.

Associated Press. 2025. In lawsuit over teen's death, judge rejects arguments that AI chatbots have free speech rights. *AP News*.

Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: From allocative to representational harms in machine learning. In *SIGCIS conference paper*.

Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakcı, Ö.; and Mariman, R. 2024. Generative AI can harm learning. *The Wharton School Research Paper*.

Bernardi, J.; Mukobi, G.; Greaves, H.; Heim, L.; and Anderljung, M. 2025. Societal Adaptation to Advanced AI. arXiv:2405.10295.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015. Online: Association for Computational Linguistics.

Bogen, M. 2024. Navigating Demographic Measurement for Fairness and Equity: AI Governance in Practice Guide. Technical report, Center for Democracy & Technology, AI Governance Lab. Contributions by Ariana Aboulafia, Kevin Bankston, Ridhi Shetty, and Amy Winecoff; Designed by Timothy Hoagland.

Bogen, M.; Rieke, A.; and Ahmed, S. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 492–500. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Conference on Neural Information Processing Systems (NeurIPS)*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258*.

Bommasani, R.; Klyman, K.; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2025. The 2024 Foundation Model Transparency Index. *Transactions on Machine Learning Research*.

Brayne, S.; and Christin, A. 2021. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*, 68(3): 608–624.

Business for Social Responsibility. 2021. Human Rights Assessments: Identifying Risks, Informing Strategy. *BSR*.

Capraro, V.; Lentsch, A.; Acemoglu, D.; Akgun, S.; Akhmedova, A.; Bilancini, E.; Bonnefon, J.-F.; Brañas-Garza, P.; Butera, L.; Douglas, K. M.; Everett, J. A. C.; Gigerenzer, G.; Greenhow, C.; Hashimoto, D. A.; Holt-Lunstad, J.; Jetten, J.; Johnson, S.; Longoni, C.; Lunn, P.; Natale, S.; Rahwan, I.; Selwyn, N.; Singh, V.; Suri, S.; Sutcliffe, J.; Tomlinson, J.; van der Linden, S.; Lange, P. A. M. V.; Wall, F.; Bavel, J. J. V.; and Viale, R. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. arXiv:2401.05377.

Cohen, L.; Hsieh, J.; Hong, C.; and Shen, J. H. 2025. Two Tickets are Better than One: Fair and Accurate Hiring Under Strategic LLM Manipulations. *arXiv preprint arXiv:2502.13221*.

Committee, N. A. I. A. 2023. Recommendation: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting. Technical report, National Institute of Standards and Technology.

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1571–1583. New York,

NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Creel, K.; and Hellman, D. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*.

Dai, J.; Gradu, P.; Raji, I. D.; and Recht, B. 2025. From Individual Experience to Collective Evidence: A Reporting-Based Framework for Identifying Systemic Harms. arXiv:2502.08166.

Daneshjou, R.; Vodrahalli, K.; Novoa, R. A.; Jenkins, M.; Liang, W.; Rotemberg, V.; Ko, J.; Swetter, S. M.; Bailey, E. E.; Gevaert, O.; Mukherjee, P.; Phung, M.; Yekrang, K.; Fong, B.; Sahasrabudhe, R.; Allerup, J. A. C.; Okata-Karigane, U.; Zou, J.; and Chiou, A. S. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32): eabq6147.

De Vries, H.; Bahdanau, D.; and Manning, C. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

DeepSeek-AI; :; Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; Gao, H.; Gao, K.; Gao, W.; Ge, R.; Guan, K.; Guo, D.; Guo, J.; Hao, G.; Hao, Z.; He, Y.; Hu, W.; Huang, P.; Li, E.; Li, G.; Li, J.; Li, Y.; Li, Y. K.; Liang, W.; Lin, F.; Liu, A. X.; Liu, B.; Liu, W.; Liu, X.; Liu, X.; Liu, Y.; Lu, H.; Lu, S.; Luo, F.; Ma, S.; Nie, X.; Pei, T.; Piao, Y.; Qiu, J.; Qu, H.; Ren, T.; Ren, Z.; Ruan, C.; Sha, Z.; Shao, Z.; Song, J.; Su, X.; Sun, J.; Sun, Y.; Tang, M.; Wang, B.; Wang, P.; Wang, S.; Wang, Y.; Wang, Y.; Wu, T.; Wu, Y.; Xie, X.; Xie, Z.; Xie, Z.; Xiong, Y.; Xu, H.; Xu, R. X.; Xu, Y.; Yang, D.; You, Y.; Yu, S.; Yu, X.; Zhang, B.; Zhang, H.; Zhang, L.; Zhang, L.; Zhang, M.; Zhang, M.; Zhang, W.; Zhang, Y.; Zhao, C.; Zhao, Y.; Zhou, S.; Zhou, S.; Zhu, Q.; and Zou, Y. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. arXiv:2401.02954.

Editors of HLR. 2024. Voluntary Commitments from Leading Artificial Intelligence Companies on July 21, 2023. *Harvard Law Review*, 137: 1282–1303. Available at https://harvardlawreview.org/print/vol-137/voluntary-commitments-from-leading-artificial-intelligence-companies-on-july-21-2023/.

Eloundou, T.; Beutel, A.; Robinson, D. G.; Gu-Lemberg, K.; Brakman, A.-L.; Mishkin, P.; Shah, M.; Weng, L.; and Kalai, A. T. 2024. First-Person Fairness in Chatbots. Technical report, OpenAI.

Eneanya, N. D.; Yang, W.; and Reese, P. P. 2019. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2): 113–114.

Epstein, R. A. 2018. The Many Sins of NEPA. *Texas A&M Law Review*, 6(1): 1–14.

European Commission. 2025. Third Draft of the General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts. Published on 11 March 2025.

European Parliament and Council. 2024. Artificial Intelligence Act: Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. https://artificialintelligenceact.eu/.

Executive Office of the President. 2020. Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/. Issued December 3, 2020.

Executive Office of the President. 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. 88 FR 75191.

Executive Office of the President. 2025. Removing Barriers to American Leadership in Artificial Intelligence. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/. 90 FR 8741.

Ferrara, E. 2024. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1).

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

Ghosh, S.; Lutz, N.; and Caliskan, A. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 463–475.

Green, B. 2022. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Computer Law & Security Review*, 45: 105681.

Green, B.; and Hu, L. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*.

Groves, L.; Metcalf, J.; Kennedy, A.; Vecchione, B.; and Strait, A. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1107–1120. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Gyevnar, B.; and Kasirzadeh, A. 2025. AI Safety for Everyone. *arXiv:2502.09288*.

Hall, P.; Mundahl, O.; and Park, S. 2025. The Pitfalls of "Security by Obscurity" and What They Mean for Transparent AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27): 28042–28051.

Handa, K.; Bent, D.; Tamkin, A.; McCain, M.; Durmus, E.; Stern, M.; Schiraldi, M.; Huang, S.; Ritchie, S.; Syverud,

S.; Jagadish, K.; Vo, M.; Bell, M.; and Ganguli, D. 2025a. Anthropic Education Report: How University Students Use Claude. Accessed: 2025-05-02.

Handa, K.; Tamkin, A.; McCain, M.; Huang, S.; Durmus, E.; Heck, S.; Mueller, J.; Hong, J.; Ritchie, S.; Belonax, T.; Troy, K. K.; Amodei, D.; Kaplan, J.; Clark, J.; and Ganguli, D. 2025b. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. *arXiv:2503.04761*.

Hasjim, B. J.; Azafar, G.; Lee, F.; Diwan, T. S.; Raju, S.; Gross, J. A.; Sidhu, A.; Ichii, H.; Krishnan, R. G.; Mamdani, M.; Sharma, D.; and Bhat, M. 2024. The AI Agent in the Room: Informing Objective Decision Making at the Transplant Selection Committee. *medRxiv*.

Hill, B. M.; and Shaw, A. 2013. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLOS ONE*, 8(6): e65782.

Ho, A., Daniel E. Xiang. 2020. Affirmative Algorithms: The Legal Grounds for Fairness as Awareness Affirmative Action at a Crossroads. *University of Chicago Law Review Online*, 2020: 134.

Hoes, E.; and Gilardi, F. 2025. Existential risk narratives about AI do not distract from its immediate harms. *Proceedings of the National Academy of Sciences*, 122(16): e2419055122.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2023. AI generates covertly racist decisions about people based on their dialect. *Nature*, 622: 482–489.

Hooker, S. 2024. On the Limitations of Compute Thresholds as a Governance Strategy. arXiv:2407.05694.

Huang, S.; Durmus, E.; McCain, M.; Handa, K.; Tamkin, A.; Hong, J.; Stern, M.; Somani, A.; Zhang, X.; and Ganguli, D. 2025. Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions. *arXiv preprint arXiv:2504.15236*.

Intelligence, I. R. C. O. A. 2024. *Challenging Systematic Prejudices: An Investigation into Gender Bias in Large Language Models*. United Nations Educational, Scientific and Cultural Organization.

Katzman, J.; Wang, A.; Scheuerman, M.; Blodgett, S. L.; Laird, K.; Wallach, H.; and Barocas, S. 2023. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14277–14285.

Kaushik, D.; Hovy, E.; and Lipton, Z. C. 2020. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. arXiv:1909.12434.

Kearns, M. 2024. Responsible AI in the generative era.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; Ma, Z.; Thrush, T.; Riedel, S.; Waseem, Z.; Stenetorp, P.; Jia, R.; Bansal, M.; Potts, C.; and Williams, A. 2021. Dynabench: Rethinking Benchmarking in NLP. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4110–4124. Online: Association for Computational Linguistics.

King, J.; Ho, D.; Gupta, A.; Wu, V.; and Webley-Brown, H. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 492–505. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Kiviat, B. 2018. The Art of Deciding with Data: Evidence from How Employers Translate Credit Reports into Hiring Decisions. *Socio-Economic Review*, 17(2): 283–309.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14): 7684–7689.

Kraft, T.; and Zheng, Y. 2021. How Supply Chain Transparency Boosts Business Value. *MIT Sloan Management Review*, 63(1): 34–40. Copyright - Copyright Massachusetts Institute of Technology, Cambridge, MA Fall 2021; Last updated - 2024-12-11; SubjectsTermNotLitGenreText - Bangladesh; United Kingdom–UK.

Kumar, A.; He, Y.; Markosyan, A. H.; Chern, B.; and Arrieta-Ibarra, I. 2025. Detecting Prefix Bias in LLM-based Reward Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 3196–3206. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

LangChain. 2025. How Klarna's AI assistant redefined customer support at scale for 85 million active users — blog.langchain.dev. https://blog.langchain.dev/customers-klarna/.

Laufer, B.; Kleinberg, J.; and Heidari, H. 2025. The Backfiring Effect of Weak AI Safety Regulation. arXiv:2503.20848.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C. A.; Manning, C. D.; Re, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R. A.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D. S.; and Zou, J. 2024. What's documented in AI? Systematic Analysis of 32K AI Model Cards. arXiv:2402.05160.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International*

Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, 3150–3158. PMLR.

Longpre, S.; Klyman, K.; Appel, R. E.; Kapoor, S.; Bommasani, R.; Sahar, M.; McGregor, S.; Ghosh, A.; Blili-Hamelin, B.; Butters, N.; Nelson, A.; Elazari, A.; Sellars, A.; Ellis, C. J.; Sherrets, D.; Song, D.; Geiger, H.; Cohen, I.; McIlvenny, L.; Srikumar, M.; Jaycox, M. M.; Anderljung, M.; Johnson, N. F.; Carlini, N.; Miailhe, N.; Marda, N.; Henderson, P.; Portnoff, R. S.; Weiss, R.; Westerhoff, V.; Jernite, Y.; Chowdhury, R.; Liang, P.; and Narayanan, A. 2025. In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI. arXiv:2503.16861.

Luccioni, S.; Gamazaychikov, B.; Hooker, S.; Pierrard, R.; Strubell, E.; Jernite, Y.; and Wu, C.-J. 2024. Light bulbs have energy ratings — so why can't AI chatbots? *Nature*, 632(8026): 736–738.

Martin, J. L.; and Wright, K. E. 2023. Bias in automatic speech recognition: The case of African American language. *Applied Linguistics*, 44(4): 613–630.

Mireshghallah, N.; Antoniak, M.; More, Y.; Choi, Y.; and Farnadi, G. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv:2407.11438.

Montoya, L. N.; Roberts, J. S.; and Hidalgo, B. S. 2025. Towards Fairness in AI for Melanoma Detection: Systemic Review and Recommendations. In *Future of Information and Communication Conference*, 320–341. Springer.

Narayanan, A.; and Kapoor, S. 2023. AI as Normal Technology. *Knight First Amendment Institute at Columbia University*. https://knightcolumbia.org/content/ai-as-normal-technology.

Narayanan, A.; and Kapoor, S. 2024. AI safety is not a model property. *AI Snake Oil*.

Newman, L. H.; and Burgess, M. 2025. Exposed DeepSeek Database Revealed Chat Prompts and Internal Data. *WIRED*.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Prasad, R.; Nguyen, L.; Schwartz, R. M.; and Makhoul, J. 2002. Automatic transcription of courtroom speech. In *INTERSPEECH*, 1745–1748.

Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2025. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *International Conference on Learning Representations (ICLR)*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *International Conference on Learning Representations (ICLR)*.

Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

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*, FAT* '20, 33–44. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

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*, AIES '22, 557–571. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Recommendation: National Artificial Intelligence Advisory Committee (NAIAC). 2024. Data Challenges and Privacy Protections for Safeguarding Civil Rights in Government. Technical report, U.S. National Artificial Intelligence Initiative Office.

Relan, P. 2025. How Generative AI Can Support Underfunded Schools in STEM Education. https://aibusiness.com/generative-ai/how-generative-ai-can-support-underfunded-schools-in-stem-education.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Seyyed-Kalantari, L.; Zhang, H.; McDermott, M. B.; Chen, I. Y.; and Ghassemi, M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12): 2176–2182.

Shen, J. H.; and Guestrin, C. 2025. Societal Impacts Research Requires Benchmarks for Creative Composition Tasks. *arXiv preprint arXiv:2504.06549*.

Snap Inc. 2023. Evolving My AI with Sponsored Links powered by Microsoft Advertising.

Surani, F.; Suzgun, M.; Raman, V.; Manning, C. D.; Henderson, P.; and Ho, D. E. 2024. AI for Scaling Legal Reform: Mapping and Redacting Racial Covenants in Santa Clara County. *Stanford RegLab Working Paper*.

Suresh, H.; and Guttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385534.

Sylvestre, S. 2025. Harnessing the Power of Generative AI to Close the Achievement Gap.

https://www.sir.advancedleadership.harvard.edu/articles/harnessing-power-generative-ai-close-achievement-gap.

Teo, C.; Abdollahzadeh, M.; and Cheung, N.-M. M. 2023. On Measuring Fairness in Generative Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 10644–10656. Curran Associates, Inc.

Tomasev, N.; McKee, K. R.; Kay, J.; and Mohamed, S. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 254–265. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Uma Rani. 2024. AI Labour Disclosure Initiative: Recognizing the social cost of human labour behind automation.

Wachter, S. 2020. Affinity Profiling and Discrimination by Association in Online Behavioral Advertising. *Berkeley Technology Law Journal*, 35(2): 367–430.

Wallace, E.; Watkins, O.; Wang, M.; Chen, K.; and Koch, C. 2025. Estimating Worst-Case Frontier Risks of Open-Weight LLMs. arXiv:2508.03153.

Wang, A.; Barocas, S.; Laird, K.; and Wallach, H. 2022. Measuring Representational Harms in Image Captioning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Wang, A.; Datta, T.; and Dickerson, J. P. 2024. Strategies for Increasing Corporate Responsible AI Prioritization. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Wang, A.; and Russakovsky, O. 2023. Overwriting Pretrained Bias with Finetuning Data. *International Conference on Computer Vision (ICCV)*.

Watkins, E. A.; and Chen, J. 2024. The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 764–775. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Wilson, K.; and Caliskan, A. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Xiang, A. 2024. Fairness & Privacy in an Age of Generative AI. *Science and Technology Law Review*, 25(2).

Zhang, A. K.; Ji, J.; Menders, C.; Dulepet, R.; Qin, T.; Wang, R. Y.; Wu, J.; Liao, K.; Li, J.; Hu, J.; Hong, S.; Demilew, N.; Murgai, S.; Tran, J.; Kacheria, N.; Ho, E.; Liu, D.; McLane, L.; Bruvik, O.; Han, D.-R.; Kim, S.; Vyas, A.; Chen, C.; Li, R.; Xu, W.; Ye, J. Z.; Choudhary, P.; Bhatia, S. M.; Sivashankar, V.; Bao, Y.; Song, D.; Boneh, D.; Ho, D. E.; and Liang, P. 2025. BountyBench: Dollar Impact of AI Agent Attackers and Defenders on Real-World Cybersecurity Systems. arXiv:2505.15216.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.