

The Hidden Tax of AI Safety: Residual Risk Redistribution in Automated Governance

Henrietta Boowuo, Ernest Ofori-Mensah

School of Applied and Creative Computing, Purdue University
West Lafayette, Indiana 47906, USA
hboowuo@purdue.edu, eforime@purdue.edu

Abstract

Current AI safety paradigms operate on a flawed assumption: a technical reduction in system-wide error equates to a universal reduction in harm. This paper challenges the "optimization logic" by revealing that current safety safeguards do not eliminate risk but rather redistribute it. We define this phenomenon as the Hidden Tax of Safety, a condition where the burdens of imperfect AI protections are systematically offloaded onto marginalized populations. Using a sociotechnical analysis of content moderation, deepfake detection, and Automated Speech Recognition (ASR), we show that "safety" measures often result in Individualized Responsibility.

When an AI safeguard fails or produces a false positive, the burden of correcting it, appealing a ban, proving identity, or overcoming linguistic exclusion falls exclusively on the user. The paper moves beyond technical robustness to propose a Fair Risk Distribution framework grounded in distributive justice. We argue that for an AI system to be truly "safe," the cost of its failures must be borne by the developer and the institution, not the vulnerable subject. We propose "Burden-Shifting Governance" mechanisms, including automated redress protocols and equity-weighted safety metrics.

Introduction

As generative artificial intelligence (GenAI) systems become increasingly embedded within education, labor, governance, communication, and public administration, AI safety has emerged as a defining challenge of contemporary technological governance (Crawford 2021a; Bommasani et al. 2021). Notably, prevailing safety paradigms remain dominated by an optimization-centered logic that evaluates system reliability primarily through aggregate technical metrics such as robustness, alignment, accuracy, and benchmark performance (National Institute of Standards and Technology 2023). Within this framework, reductions in overall system error are often treated as evidence of universally improved outcomes. This assumption obscures a more fundamental sociotechnical reality: the harms that remain after safeguards are implemented are rarely distributed equally across populations (Hoffman 2019; Suresh and Gutttag 2021).

This paper argues that contemporary AI safety infrastructures frequently operate through a form of distributive

asymmetry in which the burdens of technological imperfection are systematically transferred onto marginalized users. Safety systems, including automated content moderation, AI-generated text detectors, facial recognition technologies, and automated speech recognition (ASR), are typically designed to maximize institutional protection and operational scalability. When these systems fail, however, institutions rarely absorb the consequences directly. Instead, affected individuals must bear the cognitive, emotional, temporal, and procedural costs of remediation by appealing moderation decisions, correcting inaccurate captions, disproving algorithmic suspicion, or repeatedly verifying identity (Eubanks 2018; Herd and Moynihan 2018). We conceptualize this phenomenon as the *Hidden Tax of Safety*: the unequal redistribution of residual harm produced by imperfect AI safeguards.

Recent scholarship increasingly demonstrates that algorithmic systems reproduce and amplify existing social hierarchies rather than functioning as neutral computational infrastructures. For example, Liang et al. (2023b) shows that AI-generated text detectors disproportionately classify non-native English writing as machine-generated because second-language writing patterns often resemble the statistical regularities associated with synthetic text. Similarly, ASR systems exhibit significantly higher error rates for African American Vernacular English (AAVE), accented English varieties, and other linguistically marginalized speech communities (Koenecke et al. 2020b). Research on facial recognition has documented substantial demographic disparities, with darker-skinned individuals, particularly women, experiencing disproportionately high rates of misidentification (Buolamwini and Gebru 2018a). Taken together, these studies suggest that safeguard failures are not merely isolated technical anomalies. Once embedded into institutional infrastructures, uneven error distribution becomes a mechanism of structural inequality.

Existing approaches to AI fairness and safety struggle to account for these dynamics because they often prioritize aggregate optimization while neglecting the unequal distribution of residual harm. Hoffmann argues that fairness metrics frequently reduce justice to problems of statistical calibration while obscuring deeper institutional inequities (Hoffmann 2019). Similarly, Selbst et al. (2019) caution that computational fairness frameworks often abstract social prob-

lems into technical variables, thereby masking the structural conditions through which harms are experienced. Building on these critiques, this paper argues that AI safety itself increasingly functions as a mechanism of burden redistribution. The central issue is therefore not simply whether systems are accurate on average, but whether the remaining costs of technological failure are concentrated among groups with the least institutional capacity to contest them.

Drawing on sociotechnical theory, distributive justice, and critical AI governance scholarship, this paper reconceptualizes AI safety as a problem of residual risk distribution rather than narrow technical optimization. We introduce *Residual Risk Redistribution Theory* (RRRT). This framework explains how imperfect safeguards transfer corrective responsibility from institutions to users through what we term *correction labor*: the cognitive, temporal, and procedural effort required to overcome safeguard failures. Rather than asking only how much risk remains after optimization, RRRT asks who bears the burden of the remaining risk and under what institutional conditions.

The paper proceeds as follows. Section 2 develops the theoretical foundations of RRRT and introduces the concept of the *Safety Illusion*. Section 3 examines the mechanisms through which burdens are redistributed from institutions to users, including unequal exposure, correction labor, and institutional opacity. Section 4 synthesizes evidence across multiple AI domains, including AI detection systems, ASR, content moderation, and facial recognition, to demonstrate recurring patterns of redistributive harm. Section 5 introduces the Fair Risk Distribution Framework (FRDF), Equity-Weighted Residual Risk (EWRR), and Presumptive Humanity Principle (PHPs) as governance interventions designed to address distributive asymmetries in AI safety. Finally, Section 6 considers the broader implications of distributive AI governance for regulatory institutions and democratic legitimacy.

Finally, this paper argues that the central challenge of AI governance is no longer merely reducing technical error, but governing how unavoidable technological failures are socially distributed. As AI systems increasingly function as infrastructures of participation, recognition, and institutional access, the legitimacy of AI safety will depend not only on whether systems become more accurate, but on whether democratic institutions can prevent the burdens of technological imperfection from being systematically transferred onto those with the least power to resist them.

Residual Risk Redistribution Theory

Limits of Optimization-Centric Safety

Contemporary AI governance frameworks are increasingly organized around technical robustness (Raji et al. 2020). Across industry, policy, and research communities, safer AI systems are commonly associated with improved alignment, benchmark performance, adversarial resilience, and lower aggregate error rates (National Institute of Standards and Technology 2023). Within this paradigm, safety is primarily treated as a measurable engineering outcome. If a system demonstrates higher overall accuracy or reduced harm-

ful outputs, it is assumed to produce socially improved outcomes (Birhane 2021; Selbst et al. 2019).

While these interventions remain important, they also generate what this paper conceptualizes as the *Safety Illusion*: the assumption that reductions in aggregate system error necessarily correspond to reductions in socially distributed harm. Aggregate optimization can conceal significant subgroup disparities because harms produced by automated systems are rarely experienced uniformly across populations (Buolamwini and Gebru 2018a). A system reporting exceptionally high overall accuracy may still expose marginalized users to disproportionately high rates of misclassification, procedural burden, or exclusion. In such cases, benchmark success obscures distributive inequality by rendering unequal exposure statistically invisible (Suresh and Gutttag 2021).

This limitation becomes particularly consequential once safeguards are embedded into institutional infrastructure. AI systems no longer function merely as technical artifacts; they increasingly regulate participation, legitimacy, visibility, and access across education, employment, communication, surveillance, and digital governance. Under these conditions, even relatively small disparities in system performance can generate substantial downstream consequences (Costanza-Chock 2020; Eubanks 2018). A false positive generated by an AI-detection system can trigger reputational damage, disciplinary suspicion, or procedural burdens that vulnerable users disproportionately bear. As Selbst et al. (2019) argues, computational fairness frameworks often abstract social problems into technical variables while obscuring the institutional conditions through which harms are experienced.

The emphasis on benchmark optimization also reinforces a hidden assumption in AI governance: that the "average user" serves as a socially neutral baseline around which safety can be designed. However, AI systems do not operate within neutral environments. Training data, annotation practices, optimization objectives, and deployment contexts are all shaped by historical inequalities and institutional power structures (Noble 2018; Eubanks 2018). Crawford describes AI infrastructure not as isolated technical systems, but as political and social arrangements embedded within existing relations of power (Crawford 2021b). Consequently, safeguards optimized for dominant populations may still intensify exclusion for users positioned outside normative institutional expectations. We argue that the critical governance question is no longer simply how much risk remains after safeguards are implemented, but who bears the burden of that risk.

The Residual Risk Redistribution

To address these limitations, this paper introduces *Residual Risk Redistribution Theory* (RRRT), a sociotechnical framework explaining how imperfect AI safeguards redistribute residual harm across unequal institutional structures. RRRT departs from conventional safety paradigms by arguing that safeguards do not eliminate harm through optimization; rather, they reorganize the distribution of unavoidable harms socially, institutionally, and politically.

Stage	Technical Process	Pro-	Redistributive Outcome
Optimization	Aggregate error minimization		Minority variance obscured
Deployment	Institutional embedding		Safeguard becomes infrastructure
Failure Event	False positive/negative		Residual harm emerges
Burden Transfer	User remediation		Responsibility individualized
Institutional Shielding	Proprietary opacity		Accountability asymmetry
Stratification	Repeated unequal exposure		Structural inequality reproduced

Table 1: Residual Risk Redistribution Lifecycle

Within traditional engineering and risk management frameworks, residual risk refers to the level of risk that remains after mitigation strategies have been implemented (National Institute of Standards and Technology 2023). RRRT extends this concept by arguing that residual risk in AI systems cannot be understood purely as a technical remainder. Because automated systems interact with unequal social environments, failures in safeguarding become distributive events. False positives, moderation errors, accessibility failures, and biometric misidentifications do not affect all users equally. Instead, their burdens frequently cluster among populations with reduced institutional power, limited procedural recourse, or non-dominant forms of communication (Birhane 2021; Devinney, Miltenburg, and Sluyter-Beltman 2023). Central to RRRT is the *Residual Risk Redistribution Lifecycle*, which explains how optimization-centered safeguards evolve into mechanisms of structural inequality once deployed institutionally (see Table 1).

The lifecycle begins with optimization processes oriented toward improving aggregate performance. During this stage, subgroup disparities may remain hidden because minority variance becomes diluted within larger populations. Once deployed institutionally, safeguards transition from isolated technical systems to infrastructures that govern communication, access, enforcement, and participation (Suresh and Guttag 2021). Failures occurring within these infrastructures subsequently generate residual harms that require remediation. Rather than institutions absorbing these costs directly, responsibility is frequently transferred downward onto users through appeals, correction labor, repeated verification, or procedural navigation (Herd and Moynihan 2018).

Importantly, institutions often remain insulated from accountability through proprietary opacity, limited transparency, and asymmetrical access to technical knowledge (Pasquale 2015; Zuboff 2019). RRRT conceptualizes this condition as *accountability asymmetry*: institutions retain the efficiency gains of automation, while marginalized users bear the burdens of safeguard failures. Over time, repeated

unequal exposure to residual harms contributes to broader forms of structural stratification in which technological systems reproduce and stabilize existing inequalities (Birhane 2021; Crawford 2021b).

RRRT therefore shifts the analytical focus of AI safety away from narrow questions of technical accuracy toward broader questions of distributive legitimacy. Rather than asking whether safeguards function effectively “on average”, the framework asks how the unavoidable imperfections of automated governance are distributed across social and institutional contexts.

Distributive Governance

The theoretical foundations of RRRT draw from broader traditions of distributive justice and critical sociotechnical scholarship. Rawls (1971) argued that just institutions must be evaluated according to how they affect the least advantaged members of society rather than through aggregate utility. Applied to AI governance, this principle suggests that high overall performance cannot ethically justify systems whose residual harms are systematically concentrated among marginalized groups.

Similarly, Birhane emphasizes that algorithmic systems inherit and amplify existing relational inequalities because they are embedded within historically unequal social structures (Birhane 2021). AI infrastructures cannot be treated as neutral technical tools operating independently of institutional power. Crawford argues that AI systems function as political infrastructures governing visibility, legitimacy, and social participation (Crawford 2021b).

Taken together, these perspectives support a broader reconceptualization of AI safety as distributive governance. As automated systems increasingly mediate institutional participation, governance is no longer limited to regulating technological outputs. It increasingly involves governing the distribution of technological vulnerability across society. The central ethical challenge of AI safety is therefore not merely minimizing error but ensuring that the burdens of unavoidable technological imperfection are not systematically imposed on those with the least institutional power to resist them.

Mechanisms of Burden Transfer

The redistribution of residual harm does not occur randomly. Rather, it emerges through recurring sociotechnical mechanisms embedded within the design, deployment, and governance of AI systems. These mechanisms transform technical imperfections into institutionalized forms of unequal exposure, procedural friction, and asymmetric accountability, as shown in (see Table 2). While safeguards are frequently framed as neutral infrastructures of protection, their operational logic often privileges users whose language, behavior, and identity align most closely with dominant institutional norms. Consequently, the benefits of automation become centralized within institutions, while the burdens of technological imperfection are externalized onto vulnerable populations.

Computational Normativity

AI systems are trained on data infrastructures that inevitably encode dominant linguistic patterns, cultural assumptions, and institutional expectations. Although machine learning models are often presented as objective computational systems, their outputs remain deeply shaped by the distributions embedded within training corpora, annotation practices, benchmark construction, and optimization objectives (Birhane 2021; Crawford 2021b). What appears computationally “normal” is therefore not neutral, but historically contingent and institutionally produced.

This dynamic creates what may be understood as *normative infrastructure*: the computational privileging of dominant populations through the operationalization of majority norms as technical baselines. Linguistic systems, for instance, are commonly optimized around standardized forms of English and highly represented speech patterns. In contrast, accented speech, Indigenous languages, dialectal variation, and multilingual communication remain comparatively underrepresented (Koenecke et al. 2020b). Similarly, automated moderation systems frequently classify dominant communicative practices as legitimate while disproportionately interpreting marginalized linguistic expression as suspicious, toxic, or deviant.

Importantly, these disparities cannot be reduced to isolated dataset imbalances. They reflect broader institutional assumptions concerning whose communication is considered intelligible, trustworthy, and normatively acceptable within digital infrastructures. As Benjamin (2019a) argues, automated systems often reproduce “default discrimination” by embedding historical inequities into technological architectures that appear formally neutral. AI infrastructures, therefore, do not simply process social reality; they actively organize participation by computationally privileging users who conform most closely to dominant institutional expectations.

The Burden of Corrective Labor

One of the most significant consequences of safeguard failure is the emergence of what this paper conceptualizes as *correction labor*: the cognitive, emotional, temporal, and institutional labor required to overcome automated error. Correction labor represents the practical mechanism through which residual harm becomes individualized. Rather than institutions absorbing the costs of technological imperfection, users themselves become responsible for restoring accuracy, legitimacy, or procedural standing after a safeguard failure (Eubanks 2018; Zuboff 2019).

Correction labor manifests across multiple AI domains. Users may be required to appeal moderation decisions, manually correct inaccurate captions generated by ASR systems, complete repeated identity verification procedures, or defend themselves against algorithmically generated suspicion (Eubanks 2018; DeVito 2022). In educational settings, students incorrectly flagged by AI detection systems may need to provide drafting histories, explain their writing processes, or navigate disciplinary review procedures to disprove automated suspicions (Liang et al. 2023b). Similarly, speakers

Mechanism	Technical Framing	Redistributed Burden
Unequal Exposure	Statistical variance	Demographic exclusion
Correction Labor	User remediation	Cognitive/time burden
Institutional Shielding	Proprietary safeguards	Limited recourse
Automated Enforcement	Safety optimization	Presumption of suspicion

Table 2: Mechanisms of Redistribution

whose accents are poorly recognized by ASR systems often spend additional time and cognitive effort correcting transcription errors to achieve equivalent communicative outcomes.

Although these activities are frequently treated as minor inconveniences, their cumulative effects are structurally significant. Correction labor redistribution shifts the operational costs of AI safety away from institutions and onto individuals least able to absorb them. What institutions gain through automation efficiency and scalability is offset by the hidden remediation burden imposed on users. Technological failure, therefore, becomes a form of invisible labor disproportionately borne by marginalized populations (Star and Strauss 1999; Gray and Suri 2019).

Correction labor also extends beyond measurable time expenditure. Repeatedly contesting moderation decisions, disproving algorithmic accusations, or correcting machine-generated inaccuracies can lead to institutional fatigue and diminished procedural trust (Herd and Moynihan 2018; Vaccaro, Sandvig, and Karahalios 2021). In this sense, correction labor is not merely technical remediation; it is a governance burden produced through the unequal distribution of technological imperfection.

Institutional Burden Transfer

The broader consequence of normative infrastructure and correction labor is the emergence of *institutional burden transfer*: a governance dynamic in which institutions retain the efficiency gains of automation. At the same time, users absorb the residual harms resulting from safeguard failures. AI systems enable organizations to scale moderation, verification, surveillance, accessibility, and enforcement processes at unprecedented speed and cost (Zuboff 2019; Eubanks 2018). However, the imperfections inherent within these systems do not disappear through optimization. Instead, the burdens associated with those imperfections are displaced downward onto users through individualized remediation requirements (Herd and Moynihan 2018; Gillespie 2018).

This process produces what RRRT conceptualizes as *individualized responsibility*. Institutional legitimacy increasingly depends on automated systems; responsibility for correcting those systems’ failures is transferred to affected individuals. Users become responsible for proving author-

ship, restoring communicative clarity, verifying identity, or contesting algorithmic suspicion within systems they neither designed nor control (Pasquale 2015). Moreover, most users lack meaningful insight into how safeguards operate, how thresholds are determined, or how automated decisions may be challenged. As a result, procedural recourse becomes unevenly accessible, particularly for populations already marginalized by linguistic, racial, or institutional barriers (DeVito 2022).

Taken together, these mechanisms reveal that AI safety infrastructures operate not only as systems of protection, but also as systems of burden allocation. The central issue is therefore not simply whether safeguards reduce harm overall, but whether the social costs of unavoidable technological failure are distributed in ways that reinforce existing inequalities. AI governance increasingly operates by managing procedural friction, institutional access, and technological vulnerability, thereby making the redistribution of residual harm a central political and ethical problem in automated societies (Crawford 2021b; Herd and Moynihan 2018).

Cross-Domain Patterns of Residual Harm

Residual Risk Redistribution Theory (RRRT) becomes most visible when AI safeguards are examined across domains that appear technically distinct but share a common governance pattern. AI detection systems, automated speech recognition, content moderation, and facial recognition technologies are typically justified by different institutional objectives: academic integrity, accessibility, platform safety, and security. However, each domain reveals the same redistributive logic (see Table 3). Safeguards are introduced to reduce institutional risk, but when they fail, the burdens of correction, proof, appeal, or misidentification are disproportionately shifted onto populations already marginalized by language, race, disability, nationality, or institutional power (Costanza-Chock 2020; Birhane 2021). Residual harm is therefore not merely a by-product of technical imperfection; it is a patterned outcome of automated governance.

AI Detection and Academic Integrity

AI-generated text detection systems have become increasingly attractive to educational institutions seeking scalable solutions amid the rapid adoption of large language models (Liang et al. 2023a; Weber-Wulff et al. 2023). These systems promise to preserve academic integrity by distinguishing student-authored work from machine-generated content. However, the apparent efficiency of automated detection conceals a significant distributive problem. When AI detectors produce false positives, the institutional burden of proving misconduct is often reversed. Students become responsible for proving that their own writing is human.

This reversal disproportionately affects non-native English writers. Liang et al. demonstrate that widely used GPT detectors frequently misclassify second-language writing as AI-generated because non-native linguistic patterns often resemble the statistical regularities associated with synthetic text (Liang et al. 2023b). Writing that is more formulaic, grammatically constrained, or shaped by second-language acquisition may therefore appear computationally

“machine-like” despite being fully human-authored. Linguistic difference becomes algorithmic suspicion.

From the perspective of RRRT, AI detection systems illustrate how residual risk becomes individualized through institutional procedures. Institutions gain a scalable enforcement mechanism, while students absorb the burden of contesting opaque technical judgments. This may involve producing drafts, explaining writing processes, attending hearings, or managing the reputational consequences of suspected misconduct. In effect, students are increasingly required to *establish human authorship* before being subject to algorithmic suspicion. Importantly, this burden is not distributed equally. It falls most heavily on students whose writing diverges from dominant academic English norms and who may already occupy precarious institutional positions.

The ethical issue, therefore, extends beyond detector accuracy. Even if AI detectors improve technically, the governance structure surrounding their use may continue to reproduce distributive injustice if suspicion is automated while accountability remains individualized. AI detection thus reveals the hidden tax of safety in educational governance: institutions protect themselves against uncertainty by transferring residual procedural burdens onto students least positioned to contest them.

ASR and Linguistic Marginalization

Automated Speech Recognition (ASR) systems are often presented as accessibility safeguards across education, healthcare, employment, and digital communication (Koencke et al. 2020a; Tatman 2017). By converting speech into text, ASR promises broader participation for users who rely on captions, transcripts, or speech-based interfaces. In practice, substantial evidence demonstrates that ASR performance remains uneven across accents, dialects, and non-dominant speech varieties. Koencke et al. found significant racial disparities in commercial ASR systems, with markedly higher error rates for Black speakers than for White speakers (Koencke et al. 2020b). These disparities reveal how accessibility infrastructures can simultaneously expand participation for some users while reproducing linguistic inequality for others.

From the perspective of RRRT, the central issue is not merely unequal transcription accuracy but unequal correction burden. When ASR systems fail, users must engage in repetitive clarification, manual caption correction, transcript editing, or conversational repair to restore communicative accuracy. These activities constitute forms of corrective labor that are rarely captured in institutional assessments of automation value, as organizations typically evaluate accessibility systems based on efficiency, scalability, and cost reduction rather than the labor required to repair machine-generated errors (Gray and Suri 2019; Star and Strauss 1999).

This burden is not distributed evenly. Accented speakers, multilingual users, dialect speakers, and linguistically marginalized communities often face repeated misrecognition in environments where accurate transcription is tied to participation, visibility, and credibility (Fu, Harrison, and Lee 2023). In educational settings, inaccurate captions may

distort instructional meaning (Koencke et al. 2020a). In professional environments, they may undermine communicative legitimacy and institutional recognition. What appears institutionally as an accessibility safeguard may therefore operate experientially as an infrastructure of friction. ASR systems thus demonstrate the dual character of AI accessibility technologies. Institutions benefit from automation because it reduces labor and costs associated with providing access services at scale. However, when safeguards fail, remediation costs become individualized.

Content Moderation and Linguistic Legitimacy

Automated content moderation systems are typically justified as safeguards against hate speech, harassment, misinformation, and platform abuse (Gillespie 2018; Robers 2019). Their institutional purpose is to make digital environments safer and more governable at scale. However, moderation systems also demonstrate how dominant linguistic norms become computationally privileged forms of legitimacy.

Research on racial bias in hate speech detection illustrates this problem clearly. Sap et al. show that toxicity datasets and moderation systems frequently associate markers of African American English (AAVE) with offensive or abusive language, partly because annotators often fail to distinguish dialectal expression from harmful intent (Sap et al. 2019a). Related work demonstrates that attempts to debias moderation systems remain difficult when bias is embedded within annotation practices and dataset construction rather than solely within model architecture (Zhou et al. 2021).

The same logic extends beyond AAVE to Indigenous language use, reclaimed language, political speech, and culturally specific forms of expression. Communities may employ terms, rhetorical styles, or communicative practices that appear suspicious to automated systems trained predominantly on dominant-language corpora and decontextualized moderation labels. Reclaimed slurs, activist discourse, and local political expression may therefore be misclassified when systems lack contextual understanding (Sap et al. 2019b; DeVito, Walker, and Vance 2021). In such cases, safeguards designed to reduce harm may simultaneously suppress legitimate expression by groups already marginalized within dominant public spheres.

Through RRRT, content moderation can be understood as a system that redistributes expressive risk. Platforms gain the ability to manage content at scale and reduce reputational or regulatory exposure. At the same time, users bear the residual costs of false positives, shadowbanning, account suspensions, appeals, and self-censorship. The result is a hierarchy of linguistic legitimacy in which dominant communicative norms move through governance systems with comparatively less friction, while marginalized forms of expression encounter heightened scrutiny.

Facial Recognition and Unequal Surveillance

Facial recognition systems are commonly portrayed as security safeguards that enhance verification, fraud prevention, surveillance, and access control (Browne 2015; Gates 2011). However, research on demographic disparities repeatedly

Domain		Safety Objective	Residual Burden
AI Detection	ASR Moderation	Academic integrity	Proving authorship
		Accessibility Platform safety	Correction labor Linguistic suppression
Facial Recognition	Recog-	Security	Misidentification risk

Table 3: Cross-Domain Redistribution Patterns

demonstrates that facial recognition technologies do not perform equally across populations. Buolamwini and Gebru’s *Gender Shades* study documented substantial intersectional disparities in commercial gender classification systems, with darker-skinned women experiencing significantly higher error rates than lighter-skinned men (Buolamwini and Gebru 2018a). NIST demographic evaluations similarly found uneven performance across race, sex, and age groups depending on deployment context and model architecture (Grother, Ngan, and Hanaoka 2019).

These disparities become particularly consequential because facial recognition is often deployed within high-stakes institutional environments. For example, in identity verification and fraud detection, misidentification can trigger detention, exclusion, investigation, or reputational harm (Buolamwini and Gebru 2018b; Benjamin 2019b). The technical language of false positives and false negatives, therefore, understates the social significance of these failures. A false match can lead to an encounter with state authorities; a failed verification can result in denial of institutional access.

From the perspective of RRRT, facial recognition demonstrates how security optimization redistributes vulnerability in surveillance. Institutions benefit from expanded monitoring capacity, faster identification, and scalable verification systems, while the residual risks of misidentification are disproportionately borne by populations already exposed to heightened surveillance. This creates a form of asymmetric safety in which institutional security is enhanced at the cost of increased procedural vulnerability for marginalized groups.

Fair Risk Distribution Framework

The preceding sections demonstrated that contemporary AI safeguards frequently redistribute residual harm rather than eliminate it. If AI governance continues to evaluate safety primarily through aggregate optimization, then systems capable of reproducing substantial distributive inequities may still be institutionally classified as “safe”. Addressing this limitation requires a shift from robustness-centric governance toward frameworks that evaluate how technological burdens are socially allocated. In response, this paper proposes the *Fair Risk Distribution Framework* (FRDF): a governance architecture designed to ensure that the unavoidable imperfections of AI systems are managed through principles of distributive justice rather than displaced onto vul-

Pillar	Governance Goal	Institutional Requirement
Proportionality	Consequence-aware evaluation	Harm-weighted assessment
Shared Accountability	Reverse burden transfer	Institution-funded remediation
Accessibility Equity	Inclusive participation	Subgroup performance floors
Procedural Justice	Contestability	Human review pathways
Participatory Governance	Democratic legitimacy	Community oversight

Table 4: Fair Risk Distribution Framework

nerable populations as depicted in Table 4. The FRDF consists of five interdependent governance pillars: proportionality, shared accountability, accessibility, equity, procedural justice, and participatory governance (see Figure 1).

The FRDF does not reject technical safety measures such as robustness, alignment, or risk mitigation. Instead, it situates them within a broader sociotechnical governance framework that evaluates safety by how residual harms are distributed among populations and institutions.

The first pillar, *proportionality*, challenges the assumption that all technical errors carry equivalent ethical significance. Conventional evaluations often treat false positives and false negatives as interchangeable optimization problems (Kleinberg, Mullainathan, and Raghavan 2016; Barocas and Selbst 2016). However, the downstream consequences of these errors vary substantially across contexts and populations.

A false positive generated by an AI-detection system may expose a student to disciplinary suspicion and reputational harm, while a moderation error may suppress political expression or disproportionately silence marginalized communities. Proportionality, therefore, requires institutions to evaluate safeguards based on the severity and concentration of downstream harms rather than on aggregate accuracy.

The second pillar, *shared accountability*, addresses the redistribution of corrective burden identified throughout this paper. Current governance models frequently externalize the costs of safeguard failure onto users through appeals, repeated verification, and correction labor. Under the FRDF, institutions that deploy AI safeguards retain responsibility for the social consequences of system imperfections. This includes providing meaningful remediation pathways, accessible review mechanisms, and institutionally supported appeals rather than forcing affected individuals to absorb remediation burdens independently.

Accessibility equity, the third pillar, reframes performance disparities as problems of participatory inclusion rather than isolated technical deficiencies. Systems cannot be considered socially safe if marginalized users experience systematically reduced access, increased procedural friction, or elevated exposure to automated failure. Institutions should

therefore establish subgroup performance floors to prevent systems with substantial demographic disparities from being deployed solely based on strong aggregate performance. Accessibility, within this framework, becomes an issue of institutional recognition rather than an optional usability enhancement.

The fourth pillar, *procedural justice*, recognizes that AI safeguards increasingly function as governance infrastructures regulating participation, legitimacy, and institutional access. Users must therefore possess meaningful opportunities to challenge, contest, and understand automated decisions. Procedural justice requires transparent review pathways, human oversight mechanisms, and accessible recourse systems capable of addressing the asymmetries produced by opaque automated governance. Without contestability, safeguards risk functioning as unaccountable systems of institutional authority.

Finally, *participatory governance* addresses the democratic legitimacy of AI safety itself. Decisions about acceptable levels of technological risk cannot remain under the exclusive control of developers, corporations, or institutional administrators. Communities disproportionately affected by safeguard failures must participate in defining governance standards, remediation expectations, and safety thresholds. This principle aligns with broader traditions of participatory justice and challenges the concentration of epistemic authority within centralized technical institutions.

Equity-Weighted Residual Risk (EWRR)

A central limitation of contemporary AI safety evaluation is its dependence on aggregate performance metrics that treat residual harms as statistically neutral. Accuracy, precision, robustness, and benchmark optimization remain dominant indicators of system quality across both industry and regulatory environments (Raji et al. 2020; Thomas and Uminsky 2021). However, these measures often obscure how the remaining harms of automated systems are distributed socially. A safeguard may achieve exceptionally high average performance while still exposing particular populations to disproportionately high levels of procedural burden, exclusion, surveillance, or misidentification.

To address this limitation, this paper introduces *Equity-Weighted Residual Risk* (EWRR), a distributive governance metric for evaluating the social concentration of residual technological harm. EWRR extends beyond conventional safety assessment by asking not merely how much risk remains after optimization, but where that remaining risk is concentrated and who bears its consequences.

The central principle underlying EWRR is straightforward: a system cannot be considered socially safe if its residual harms cluster systematically within vulnerable populations. Concentrated exposure to safeguard failure constitutes a governance problem even when aggregate metrics appear favorable. A facial recognition system that achieves high overall accuracy while disproportionately misidentifying darker-skinned users, cannot be ethically legitimized through average performance. Similarly, moderation systems that disproportionately suppress marginalized linguistic communities remain distributively unsafe regardless of

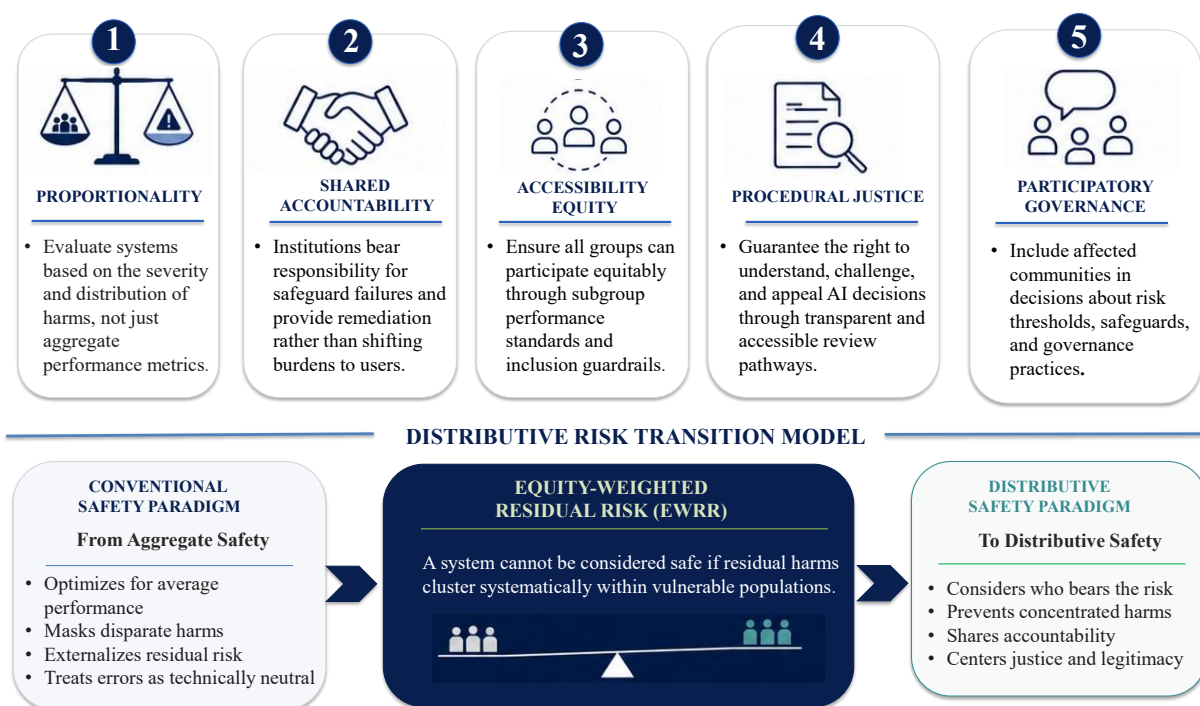


Figure 1: Residual Risk Redistribution Theory (RRRT) framework illustrating how optimization-centered AI safeguards redistribute residual harms through sociotechnical infrastructures. The framework traces the progression from aggregate optimization and institutional deployment to burden transfer, accountability asymmetry, and structural stratification. The lower governance layer presents the proposed interventions: the Fair Risk Distribution Framework (FRDF), Equity-Weighted Residual Risk (EWRR), and Presumptive Humanity Principles (PHPs), designed to mitigate the unequal distribution of technological vulnerability.

benchmark success.

EWRR therefore reframes residual error as a sociotechnical distribution problem rather than a purely computational remainder. Instead of evaluating safeguards solely by total error reduction, institutions must assess how technological imperfections interact with preexisting inequalities related to race, language, disability, citizenship, or institutional power. In this sense, EWRR functions as a distributive complement to existing technical safety metrics by establishing a normative threshold: when residual harms become systematically concentrated within groups already exposed to structural vulnerability, additional accountability obligations emerge. These obligations may include subgroup auditing, enhanced remediation pathways, contextual deployment restrictions, or reconsideration of deployment in high-risk environments. The legitimacy of AI systems, therefore, depends not only on minimizing total harm but on preventing the unequal concentration of unavoidable harms.

Presumptive Humanity Principle

One of the most consequential features of contemporary AI governance is the normalization of automated suspicion. Across educational systems, content moderation platforms, fraud detection systems, and biometric verification infras-

tructures, individuals are increasingly required to establish legitimacy amid algorithmically generated doubt (Citron and Pasquale 2014; Benjamin 2019b). This paper proposes *Presumptive Humanity Principle* (PHPs) as a governance mechanism designed to reverse this burden structure.

The core principle of PHPs is that users should not bear the burden of disproving algorithmic suspicion. Institutions deploying safeguards must instead assume responsibility for validating automated claims before imposing disciplinary, procedural, or exclusionary consequences. This principle represents a significant departure from prevailing governance models in which algorithmic outputs are routinely treated as presumptively credible despite the well-documented limitations of automated systems.

Within educational environments, PHPs would prohibit AI-detection systems from functioning as standalone evidence of academic misconduct. Institutions would instead be required to contextualize automated outputs within broader human-centered review procedures before initiating disciplinary action. Students would not be compelled to produce extensive drafting histories or procedural evidence merely to overcome probabilistic suspicion generated by opaque systems.

In content moderation systems, PHPs would similarly re-

strict punitive automation by requiring meaningful review opportunities before sanctions such as account suspension, content removal, or visibility suppression are finalized. Because moderation systems frequently struggle with dialectal variation, reclaimed language, and culturally specific forms of expression, automated enforcement can disproportionately burden marginalized communities (Sap et al. 2019a). PHPs therefore function as safeguards against procedural overreach by ensuring that users retain meaningful opportunities to contest automated judgments before punitive action is imposed.

The relevance of PHPs is equally significant within identity verification and biometric governance systems. Facial recognition infrastructures increasingly mediate access to banking systems, employment platforms, immigration checkpoints, and public services. However, demographic disparities in biometric performance expose certain populations to disproportionately high rates of verification failure and misidentification (Grother, Ngan, and Hanaoka 2019). Under current governance arrangements, individuals subjected to failed verification often bear the burden of navigating opaque remediation procedures, despite having no control over the systems that generate those failures. PHPs would require institutions to provide accessible human-review pathways and alternative verification mechanisms before making exclusionary decisions.

Governance Implications

Beyond Technical Safety

Current regulatory frameworks increasingly acknowledge that AI systems pose societal risks that extend beyond technical malfunction (European Parliament and Council of the European Union 2024; Veale and Zuiderveen Borgesius 2021). However, most governance models continue to conceptualize safety primarily through performance, reliability, and risk classification rather than through distributive consequence. The European Union AI Act, for example, represents one of the most significant attempts to regulate high-risk AI systems through obligations related to transparency, human oversight, accountability, and risk management (European Parliament 2024). While the framework marks a substantial regulatory advance, its underlying logic remains largely centered on preventing harmful outputs rather than evaluating how residual harms become socially concentrated after deployment.

Similarly, UNESCO's Recommendation on the Ethics of Artificial Intelligence emphasizes fairness, inclusion, human rights, and non-discrimination as foundational principles of AI governance (UNESCO 2021). However, operationalizing these commitments requires moving beyond generalized ethical principles toward governance mechanisms capable of identifying and addressing redistributive harms. A safeguard may formally satisfy transparency or accountability requirements while still systematically shifting procedural burdens onto marginalized populations through correction labor, asymmetrical recourse systems, or concentrated exposure to false positives.

The central governance challenge is therefore no longer

simply whether systems generate harmful outputs, but whether institutions justly distribute the burdens of unavoidable technological imperfection. Future governance frameworks must evaluate AI systems not only by aggregate performance, but also by who bears the costs of residual failure. AI safety must consequently be reframed from a narrow engineering problem into a broader sociotechnical problem concerning institutional legitimacy, democratic accountability, and distributive justice.

Residual Risk Auditing

To operationalize distributive governance, this paper proposes adopting *Residual Risk Auditing* as an institutional and regulatory oversight mechanism. Existing AI audits frequently emphasize robustness testing, adversarial evaluation, or compliance verification (Raji et al. 2020; Falco et al. 2021). While important, these approaches remain insufficient for identifying how residual harms become concentrated among vulnerable populations. Residual Risk Auditing extends beyond technical reliability by evaluating how safeguard failures interact with social inequality, procedural burden, and institutional asymmetry.

A central component of this approach involves *subgroup disparity audits* that measure not only aggregate error rates, but also the unequal distribution of false positives, accessibility failures, moderation actions, and misidentification events across demographic and linguistic groups. Systems that disproportionately harm particular populations should trigger enhanced oversight obligations, deployment restrictions, or institutional remediation requirements. This shifts auditing away from abstract performance evaluation toward assessments of distributive consequence.

Residual Risk Auditing should also incorporate *correction labor reporting*. Institutions deploying safeguards rarely account for the cognitive effort, emotional stress, temporal burden, and procedural navigation required for users to overcome automated failures (Herd and Moynihan 2018; Alvarado, Fox, and Bernstein 2020). However, these burdens represent one of the primary mechanisms through which residual harm becomes individualized. Reporting frameworks should therefore evaluate the remediation burdens associated with appeals, caption corrections, identity verification, moderation disputes, and other user-driven recovery efforts following safeguard failures.

In addition, *recourse transparency* must become a central requirement of AI governance. Users frequently encounter opaque safeguards that lack meaningful information about how decisions are generated, challenged, or reversed (Kaminski 2019; Citron and Pasquale 2014). Institutions should therefore disclose available review pathways, escalation procedures, subgroup risk profiles, and remediation mechanisms associated with deployed safeguards. Within this framework, transparency is not merely informational disclosure; it is a condition of procedural legitimacy.

Finally, Residual Risk Auditing requires explicit *burden distribution analysis*. Institutions should evaluate whether automation efficiency is achieved by displacing residual harms onto populations with reduced institutional power. Systems that systematically externalize corrective burdens

onto marginalized users cannot be considered socially safe regardless of aggregate optimization performance. Auditing, therefore, becomes not merely a technical compliance mechanism, but a form of distributive governance oversight. Potential implementation mechanisms include mandatory subgroup auditing requirements, independent oversight bodies, public transparency reporting, and procedural review standards for high-risk AI deployments. Together, these mechanisms would help shift AI governance away from narrow performance compliance toward sustained oversight of how residual harms are distributed across populations.

Democratic Legitimacy

The redistributive dynamics identified throughout this paper raise a broader question of democratic legitimacy. Decisions concerning acceptable technological risk are rarely neutral technical judgments. They are political decisions concerning whose vulnerability is considered tolerable, whose burdens are rendered visible, and whose experiences are prioritized within governance systems (Winner 1980; D'Ignazio and Klein 2020). Notably, these decisions are frequently made by developers, corporations, regulators, or institutional administrators without meaningful participation from the communities most affected by safeguard failures (Costanza-Chock 2020; Abebe et al. 2020).

This asymmetry is particularly problematic because residual harms tend to cluster among populations already positioned at the margins of institutional power. Linguistically marginalized users, racial minorities, immigrants, disabled communities, and economically vulnerable populations often experience the highest levels of procedural friction, surveillance exposure, and correction labor while having the least influence over the systems governing them (Eubanks 2018; Benjamin 2019b). Democratic legitimacy, therefore, requires moving beyond top-down governance structures toward participatory models that incorporate affected communities directly into risk-definition processes.

The principle of "Nothing About Us Without Us" is especially relevant in this context. Originally associated with disability justice and participatory rights movements, the principle holds that populations subject to institutional governance must meaningfully participate in decisions affecting their lives (Charlton 1998; Costanza-Chock 2020). Applied to AI governance, this means that communities disproportionately affected by automated safeguards should help define acceptable risk thresholds, remediation standards, deployment conditions, and oversight mechanisms. Safety cannot be legitimately imposed through purely technocratic processes when the burdens of residual harm remain socially unequal.

Participatory governance also serves an epistemic function. Marginalized communities frequently possess situated forms of knowledge concerning safeguard failure that remain invisible within institutional optimization frameworks. Users experiencing repeated misrecognition, procedural burden, or moderation suppression often identify governance failures long before institutions formally acknowledge them (Haraway 1988; D'Ignazio and Klein 2020). Incorporating these perspectives into governance structures, therefore, im-

proves not only democratic legitimacy but also institutional capacity to recognize and address redistributive harms.

Conclusion

Contemporary AI governance has largely approached safety as a technical optimization problem centered on improving robustness, alignment, and overall accuracy. While these objectives remain important, they are insufficient for understanding the broader sociotechnical consequences of automated systems. AI safeguards operate within institutional environments shaped by unequal distributions of linguistic legitimacy, procedural access, political power, and technological vulnerability. As a result, the remaining harms of imperfect safeguards are rarely distributed evenly. Instead, they are often transferred onto populations already positioned at the margins of institutional protection.

This paper reframes AI safety as a problem of distributive governance rather than narrow technical containment. Through Residual Risk Redistribution Theory (RRRT), it showed how safeguards designed to reduce institutional risk can externalize the costs of imperfection through correction labor, procedural friction, automated suspicion, and unequal exposure to failure. As AI increasingly mediates participation across education, communication, employment, surveillance, and public services, governance is becoming less about regulating outputs and more about governing how technological vulnerability is distributed across society.

This shift carries important implications for AI ethics and regulation. Aggregate optimization cannot serve as the sole measure of legitimate governance if residual harms remain concentrated among populations with the least institutional power to contest them. Safety metrics that obscure subgroup burden and governance structures that externalize corrective responsibility risk transforming safeguards into infrastructures of structural inequality.

The Fair Risk Distribution Framework (FRDF), Equity-Weighted Residual Risk (EWRR), and Presumptive Humanity Principle (PHPs) proposed in this paper collectively argue for governance models centered on accountability, procedural dignity, accessibility, equity, and democratic legitimacy. These frameworks do not reject technological innovation; rather, they insist that innovation cannot be separated from the institutional distribution of its failures. A system cannot be considered socially safe simply because it reduces aggregate harm if vulnerable populations disproportionately bear the remaining burdens.

AI governance is therefore increasingly becoming governance over the distribution of technological vulnerability. The future of AI safety will depend not only on whether systems become more accurate but also on whether democratic institutions can prevent the failures of automation from being systematically imposed on those with the least power to resist them. True safety is not achieved when residual harm disappears, but when the burdens of technological imperfection are governed justly.

References

- Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Spiegel, D. G. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260.
- Alvarado, G.; Fox, S.; and Bernstein, M. S. 2020. “I can’t answer that”: Understanding the emotional and cognitive labor of algorithmic failure in everyday interactions. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1145–1158.
- Barocas, S.; and Selbst, A. D. 2016. Big data’s disparate impact. *California Law Review*, 104(3): 671–732.
- Benjamin, R. 2019a. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
- Benjamin, R. 2019b. *Race after technology: Abolitionist tools for the new jim code*. Social Forces.
- Birhane, A. 2021. Algorithmic Injustice: A Relational Ethics Approach. *Patterns*, 2(2): 100205.
- Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Browne, S. 2015. *Dark matters: On the surveillance of blackness*. Duke University Press.
- Buolamwini, J.; and Gebru, T. 2018a. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, volume 81, 77–91.
- Buolamwini, J.; and Gebru, T. 2018b. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.
- Charlton, J. I. 1998. *Nothing about us without us: Disability oppression and empowerment*. University of California Press.
- Citron, D. K.; and Pasquale, F. 2014. The scored society: Due process for automated predictions. *Washington and Lee Law Review*, 89: 1–33.
- Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. MIT Press.
- Crawford, K. 2021a. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Crawford, K. 2021b. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Devinney, H.; Miltenburg, E. v.; and Sluyter-Beltman, K. 2023. Theories of justice in algorithmic fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1024–1035.
- DeVito, M. A. 2022. From folk theories to folk algorithmic resistance: How marginalized users navigate algorithmic architectures. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–33.
- DeVito, M. A.; Walker, A. M.; and Vance, J. 2021. “Too Loud, Too Queer, Too Black”: The Impacts of Algorithmic Content Moderation on Marginalized Creative Workers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–28.
- D’Ignazio, C.; and Klein, L. F. 2020. *Data feminism*. MIT Press.
- Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- European Parliament. 2024. Artificial Intelligence Act. European Union Regulation.
- European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Falco, G.; Shneiderman, B.; Badger, J.; Bowman, R.; Henderson, P.; Hosseinmardi, H.; Kroll, J. A.; Mitchell, M.; Raji, I. D.; Seibold, C.; et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7): 566–571.
- Fu, R.; Harrison, R.; and Lee, J. 2023. AI and the Accent Gap: Evaluating Perceptual and Acoustic Biases in Speech Tools. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 412–425.
- Gates, K. 2011. *Our biometric future: Facial recognition technology and the culture of surveillance*. NYU Press.
- Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gray, M. L.; and Suri, S. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.
- Grother, P.; Ngan, M.; and Hanaoka, K. 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Technical Report NISTIR 8280, National Institute of Standards and Technology.
- Haraway, D. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3): 575–599.
- Herd, P.; and Moynihan, D. P. 2018. *Administrative burden: Policymaking by other means*. Russell Sage Foundation.
- Hoffman, A. L. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.
- Hoffmann, A. L. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society*, 22(7): 900–915.
- Kaminski, M. E. 2019. The right to an explanation, explained. *Berkeley Technology Law Journal*, 34(1): 189–254.
- Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*.

- Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, A.; Mengesha, Z.; Brooks, C.; Haynes, H.; Zuckerberg, B.; Purnell, H.; et al. 2020a. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689.
- Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020b. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689.
- Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; and Zou, J. 2023a. GPT detectors are biased against non-native English writers. *Patterns*, 4(7): 100779.
- Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; Zou, J.; and Zou, J. Y. 2023b. GPT Detectors Are Biased Against Non-Native English Writers. *Patterns*, 4(7): 100779.
- National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, U.S. Department of Commerce.
- Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
- Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Raji, I. D.; Smart, A.; White, R.; 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.
- Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.
- Roberts, S. T. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019a. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019b. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.
- 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*, 59–68. ACM.
- Star, S. L.; and Strauss, A. 1999. Layers of silence, arenas of voice: The co-construction of social worlds and information systems. *Computer Supported Cooperative Work (CSCW)*, 8(1-2): 9–30.
- Suresh, H.; and Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning lifecycle. In *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.
- Tatman, R. 2017. Gender and dialect bias in commercial speech recognition systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 73–83.
- Thomas, R.; and Uminsky, D. 2021. Reconfiguring diversity and inclusion for AI ethics. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 45–54.
- UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence.
- Vaccaro, K.; Sandvig, C.; and Karahalios, K. 2021. Contesting the algorithm: The structural challenges of appealing algorithmic decisions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
- Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the First Comprehensive Act on AI. *Computer Law & Security Review*, 42: 105582.
- Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šamoila, S.; and Špátílová, L.-M. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1): 26.
- Winner, L. 1980. Do artifacts have politics? *Daedalus*, 109(1): 121–136.
- Zhou, X.; Sap, M.; Swayamdipta, S.; Choi, Y.; and Smith, N. A. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3143–3155.
- Zuboff, S. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.