What is Safety? Corporate Discourse, Power, and the Politics of Generative AI Safety

Ankolika De*
College of Information Sciences and Technology
Pennsylvania State University
State College, Pennsylvania, USA
apd5873@psu.edu

Gabriel Lima
Max Planck Institute for Security and Privacy
Bochum, Germany
gabriel.lima@mpi-sp.org

Yixin Zou
Max Planck Institute for Security and Privacy
Bochum, Germany
yixin.zou@mpi-sp.org

Abstract

This work examines how leading generative artificial intelligence companies construct and communicate the concept of "safety" through public-facing documents. Drawing on critical discourse analysis, we analyze a corpus of *corporate safety-related statements* to explicate how authority, responsibility, and legitimacy are discursively established. These discursive strategies consolidate legitimacy for corporate actors, normalize safety as an experimental and anticipatory practice, and push a perceived participatory agenda toward safe technologies. We argue that uncritical uptake of these discourses risks reproducing corporate priorities and constraining alternative approaches to governance and design. The contribution of this work is twofold: first, to situate safety as a sociotechnical discourse that warrants critical examination; second, to caution human-computer interaction scholars against legitimizing corporate framings, instead foregrounding accountability, equity, and justice. By interrogating safety discourses as artifacts of power, this paper advances a critical agenda for human-computer interaction scholarship on artificial intelligence.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Philosophical/theoretical foundations of artificial intelligence.*

Keywords

Safety, Generative AI, Corporate Discourses, Authority, Power, Metaphors, Chatbots

ACM Reference Format:
Ankolika De, Gabriel Lima, and Yixin Zou. 2026. What is Safety? Corporate Discourse, Power, and the Politics of Generative AI Safety. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3791632

*This work was conducted while the author was a visiting PhD Scholar at the Max Planck Institute for Security and Privacy.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791632

1 Introduction

> *AI is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire.*
> — Sundar Pichai, CEO, Google [19]

Artificial Intelligence (AI) technologies are indeed expanding rapidly in capability, availability, and integration across domains such as healthcare, education, finance, and creative industries, yet their impacts and trajectories remain far more complex and uncertain than such sweeping comparisons suggest. Instead, such analogies exemplify corporate discursive tendencies to situate AI as a transformative and universal force, one whose trajectory appears both natural and unavoidable.

Our work focuses on companies that have developed and deployed general-purpose chatbot-based Generative AI (GenAI), which has been widely adopted and appropriated by various actors, as it allows for *conversations on a wide range of topics* [131]. These applications and their rapid, ongoing deployment have fueled debates about *safety*, especially as real-world incidents underscore both the potential benefits and risks of AI deployment. Recent cases—including misinformation amplification, algorithmic bias, and psychological harms to individuals [40, 60, 98, 150]—underscore that safety is not only a technical concern but also a social and political one. Safety is a contested, context-dependent concept shaped by those with the power to define it [34], encompassing technical reliability, risk mitigation, but also social, ethical, and regulatory dimensions that determine who benefits from the technology and who is harmed.

In the absence of binding, globally enforceable regulation, companies developing GenAI often position themselves as the primary arbiters of safe deployment [68]. Corporate safety narratives thus frequently serve as a de facto framework for policymakers, researchers, and the public [68]. For example, OpenAI CEO Sam Altman recently testified that *"the benefits of the tools we have deployed so far vastly outweigh the risks"* [64] as he urged the US Congress to regulate AI. Similar rhetoric exists in industry-developed safety standards, policy white papers, and funding of research and advisory institutions focused on AI risks [3]. While these initiatives indicate a commitment to safety, they also reveal a central tension of self-regulation: the same actors driving GenAI's rapid proliferation are simultaneously defining its limits and acceptable practices.

Corporate framing also shapes the AI safety research community, which frames its work as urgent and morally imperative. Yet, its narrow focus and deference to corporate interests risk limit the field's rigor, inclusivity, and societal relevance, as responses

to these tensions shape the trajectory of AI safety research [3]. In this context, safety becomes less a collective obligation and more a strategic discourse shaped by actors with the resources to steer public understanding and regulatory agendas [90, 96, 126]. Analyzing corporate narratives can therefore reveal how safety is constructed and why it is framed in particular ways.

Ahmed et al. [3] argue that institutions do not merely describe conditions but actively orient subjects toward particular norms, futures, and attachments by repeatedly aligning them with feelings of comfort, reassurance, and legitimacy. Building on this, our work examines how "safety" is discursively constructed and mobilized in corporate communications. Drawing on feminist literature that views safety as dynamic and intersubjective—shaped by inclusion, care, transparency, accountability, epistemic uncertainty, and structural injustice [96, 126]—we examine how safety is positioned to reveal the power relations embedded in GenAI governance [52, 90].

This paper asks: *How do GenAI companies construct the notion of safety in their public communications?* Using critical discourse analyses of safety-related documents (n=75) from three major GenAI companies, we examine both what these texts include and how they frame safety, focusing on the power dynamics and political implications embedded in the language.

Across company documents, we find that responsibility is widely distributed across companies, users, and governments while accountability remains unclear; governance is framed through internal committees and selective calls for regulation; risk is presented as expansive and urgent, requiring continuous monitoring, red-teaming, and iterative deployment. These discourses position companies as proactive stewards of AI while simultaneously limiting external oversight. By treating corporate communication itself as a key site of analysis, our work contributes to and extends HCI research on AI literacy and governance, providing a critical, discursive lens through which potential avenues for HCI intervention may be identified.

2 Background

Before examining AI safety, it is important to discuss the broader concept of safety and what it means to be safe, both online and offline. In psychology, safety is the second level in Maslow's hierarchy of needs, representing the human desire for security, stability, and protection from harm; some aspects of safety in this context include physical safety, health, well-being, and financial security [86]. While Maslow discussed *safety* and *security* together, Strohmayer et al. [127] drew a distinction: *"Security is defined as the protection from deliberate threats (such as an adversary) while safety as the condition of being protected from situations that are likely to cause harm (such as toxic workplace environments) [..] Put another way, securing known harms may not produce the protection necessary to keep a person safe from bodily, emotional, and psychological harm."*

Prior work has documented how groups already marginalized by discrimination or oppression, such as sex workers [89, 125], gig workers [116], and LGBTQ+ communities [88, 119], experience heightened safety harms online and offline. In particular, criminologists Harris and Woodlock coined the term safety work to describe the mental and physical labor required to remain safe, such as when survivors of domestic violence navigate digital coercive control [54]. A feminist orientation to safety, therefore, centers the needs of marginalized communities, emphasizing that technical fixes alone are insufficient; safety must also confront power structures across wider social ecologies [127]. Building on Coles-Kemp et al. [21], post-digital perspectives similarly highlight how safety and harm are experienced across entangled digital and non-digital domains, unfolding within individuals, groups, networks, and platforms [18, 116, 121]. Together, these approaches underscore the need for a holistic understanding of safety that integrates technical, social, and political dimensions.

3 Related Work

3.1 Conceptualizing AI Safety

The primary concern in AI safety lies in how current and future AI models pose risks to individuals, society, and humanity as a whole, particularly as generative AI (GenAI)-based chatbots become widely adopted and repurposed. Ferri and Gloerich [43] differentiate between AI safety work aimed at mitigating harms that existing AI models pose to society and research addressing the long-term risks associated with the potential emergence of artificial general intelligence (AGI)- human-level models that could produce existential, catastrophic outcomes [92]. Prior work has proposed numerous frameworks and taxonomies to distinguish between different AI safety risks and relevant mitigation strategies. For instance, Abercrombie et al. [2] and Shelby et al. [122] have identified several types of risks in current AI systems, including representational and allocative harms. Rauh et al. [112] highlighted gaps in existing mitigation strategies, such as modality limitations (lack of evaluation for non-text modalities), risk coverage (insufficient attention to ethical and social risks), and contextual understanding (failure to capture the broader contexts in which AI systems operate). From a policy perspective, the United Nations has similarly identified ways in which AI can pose risks to human rights [133].

Extending the distinction between current and long-term risks, Weidinger et al. [143, 144] developed taxonomies for evaluating both immediate and systemic harms from large language models, while Uuk et al. [134] and Hagendorff [51] focused on systemic risks and behaviors such as deceptive actions, power-seeking, and self-replication. Critics have argued that the emphasis on long-term, existential risks can distract from more immediate harms, including misinformation [65], perpetuation of existing inequalities [73, 149], and environmental impacts [57]. For instance, Coldicutt [20] has shown how catastrophic framings draw on cultural myths and superhero allegories to dramatize AI risk, while Helfrich [56] has argued these metaphors deflect scrutiny from immediate harms. Lazar and Nelson [78] have criticized this long-term framing, calling instead for sociotechnical approaches that foreground accountability, inclusion, and equity.

While technical measures remain important, most AI safety work has traditionally been framed as a technical problem [35]. This narrow framing can obscure the social and political dimensions of safety, particularly given scholars' arguments that safety is inherently sociotechnical [5]. In this light, AI safety discourses are significant because discourses themselves constitute and exercise power [94]. In particular, general-purpose AI companies produce

and circulate their own technical narratives of safety through social media and other channels, shaping research priorities [104]. For instance, OpenAI documented a Teen AI Literacy Blueprint, emphasizing literacy as a key component of promoting AI safety [102]. Building on this perspective, our paper examines how mainstream GenAI companies construct, articulate, and circulate particular understandings of safety, and how these framings define what is considered legitimate, necessary, and actionable.

Taking a complementary approach, Ahmed et al. [3] examined the emergence of AI safety as an epistemic field, focusing on how authority, legitimacy, and research priorities are shaped through academic institutions, funding streams, student groups, and competitions. Their analysis emphasized the field-building dynamics of AI safety and the risk of epistemic monoculture. In contrast, our work shifts attention from institutional organization to the discursive outputs of the most influential actors. We analyze how mainstream GenAI companies publicly articulate and frame "AI safety" in their materials, highlighting how these narratives construct particular ideas about the same.

3.2 The Sociotechnicality of GenAI

Extensive work has examined how to mitigate AI risks through several technical interventions, such as benchmarking [35] and different ways of training AI [95, 107]. Prior work demonstrates how relying solely on technical solutions for safety concerns is insufficient to address real-world concerns [50, 61, 73, 79, 111, 149]. The biggest concern is that technical interventions largely fail to acknowledge that AI technologies are embedded in social contexts [120], often perpetuating existing societal harms and creating new risks. Framing these safety risks as purely technical problems can deflect accountability away from those developing GenAI [30].

A sociotechnical approach to AI acknowledges that the technology is mutually shaped by social actors and institutions [13, 32, 73, 130]. Thus, any attempts to mitigate these risks must account for the fact that AI is embedded in society and existing power structures [72, 105, 120, 149], while seeking to create *"spaces for the wider participation [...] to steer our technological future on the basis of equal concern and common humanity"* [78]. To model these safety risks in a way that accounts for social structures, researchers have emphasized the need for interdisciplinary collaboration in AI research and development [120, 130]. Going beyond academic circles, Edenberg and Wood [30] have called for more participatory design efforts, including those who are disproportionally impacted by AI technologies, as a way to balance technological affordance and the social needs of users [31, 106, 129].

GenAI is also a sociotechnical artifact; its development and interpretation are shaped by broader cultural, political, and economic forces. These forces inform sociotechnical imaginaries [84, 140]—namely, shared, socially supported visions of desirable futures [29]—that frame GenAI either as a transformative tool for progress or as a source of existential societal risk [139]. For instance, Wang et al. [140] has examined how national narratives about GenAI project societal priorities and anxieties onto technology. Another way in which GenAI represents a complex sociotechnical phenomenon is through its platformization. Similar to research on social media [100, 108], GenAI can go beyond its original functions to shape broader social, cultural, and economic practices. Scholars have argued that GenAI platforms are inherently transactional and marginalizing [14]. By relying on internal, iterative, and proprietary processes to manage AI safety risks [46], GenAI concentrates power in the hands of the few who own, design, and produce these systems while motivated by capitalistic interests [68, 137]. Likewise, when discussing and talking about safety in these scenarios, taking a sociotechnical lens becomes important [9].

The acknowledgment of GenAI safety as a sociotechnical problem is also present in HCI literature. Recent work has surveyed the ways in which humans can interact with GenAI [84, 123] and explored how to design AI in a way that accounts for those who use GenAI to promote appropriate trust [69, 70, 93]. A workshop held at CHI 2025 also explored how to address AI governance through a sociotechnical lens [41]. Extending this perspective, prior work in HCI has explored toolkits as a way to understand and explore AI systems [58, 74, 147]. Approaches like Human-Centered Explainable AI (HCXAI) exemplify these perspectives by highlighting how mismatches between designer intentions and user perceptions can lead to social misattributions, thereby reinforcing the importance of accounting for social and organizational context in AI safety [42].

In this paper, we adopt a sociotechnical lens on GenAI, analyzing corporate safety narratives to examine how power, authority, and accountability are constructed. Our discourse analysis reveals how safety is framed, which harms are emphasized or downplayed, and how participation and governance are rhetorically portrayed.

3.3 Studying Discourses

In this paper, we critically examine the *discourses* employed by GenAI platforms. Discourses are significant because they operate across multiple contexts, shaping meaning in ways that allow ideas to materialize as objects [53]. Discourses are performative, shaping how knowledge and ideas are constructed, communicated, and normalized [77, 114, 124, 135]. In technology contexts, the dominant discourses reflect existing power structures and influence whose perspectives are valued [67, 99]. Repeated narratives contribute to the stabilization of knowledge, underscoring the importance of critically examining corporate and scholarly texts to understand how authority, legitimacy, and norms are produced [62].

Discourses of safety have been studied in the context of social media, during the period when companies were expanding, developing tools and features, and implementing ideals that were discursively critiqued as unsafe. Consequently, social media companies began producing public-facing safety pages and resources meant to communicate their commitment to user safety [10]. Since then, scholars from various disciplinary perspectives and epistemologies have analyzed these pages, finding that the concept of *"safety"* is often entangled with, and obscured by, capitalistic logics of surveillance, violence, and profiteering [24, 47]. Similarly, prior work has also shown how corporate discourses around privacy align with business logics, producing deterministic claims about how privacy should be defined and enforced [91]. Research has further highlighted work on the discursive mediation of tensions between free speech and safety, often drawing on similar language but yielding divergent policies [45]. Across both current and past social media platforms, such discourses normalize evolving practices,

while the labor of being and working on these systems remains precarious due to the persistent absence of standard regulations around safety and care [80].

We build upon this scholarship by examining the discursive techniques that GenAI companies employ in their public-facing communications. We adopt a broad, context-sensitive understanding of safety following Stardust et al. [126] and conceptualize it not merely as the absence of harm, danger, or risk but as the intertwining of inclusion, equitable access, transparency, and accountability. This approach recognizes that safety is socially and politically situated, shaped by structural and systemic factors, including the practices and policies of platforms [8]. By framing AI safety in this way, we emphasize the importance of its sociotechnicality.

We focus on GenAI discourses due to the recognition that evolving societal narratives around AI are shaped by cultural, political, and ethical assumptions, reflecting and obscuring the complex, value-laden, and power-infused nature of AI technologies [113]. More specifically for AI safety, the power of discourses is evident in debates concerning how long-term, existential narratives of risk have the potential to divert attention away from more pressing—and most importantly, existing—harms posed by AI [3, 20, 56, 57]. One study exploring these narratives showed that regulatory and industry actors often prioritize narrow technical fixes (such as data bias or transparency), while sidelining deeper concerns about systemic safety, power asymmetries, and the political implications of AI deployment [38].

We situate our research within critiques of Big Tech's influence, noting that public AI discourse is often shaped by corporate narratives [23, 97, 145]. Ahmed et al. [3] highlight that authority in AI safety is shaped through institutional and discursive practices such as publications, labs, funding, and prizes, which prioritize certain risks and methods while marginalizing others. We extend this perspective to corporate contexts, examining how companies use discourse to assert authority and influence perceptions of legitimate AI safety concerns. Our work contributes to the critical HCI community interested in the sensemaking, design, policy, and governance of AI. By centering discourses—which are key to understanding how power and legitimacy shape ideas that become normative in innovations—our work highlights the sociotechnical imaginaries that guide GenAI's development and deployment.

4 Methods

As one of the first empirical studies examining how safety is discursively constructed in GenAI platforms by corporate actors, our analysis focused on a small but diverse set of AI companies offering chatbot-based tools. We chose chatbots' parent companies because of their broad conversational utility rather than domain-specific functions [131]. Our goal was to develop a formative understanding of how safety is publicly framed across widely used GenAI platforms and to analyze discourses situated in real-world contexts, where users might seek information about AI safety.

We employed critical discourse analyses (CDA) as our analytical approach, grounded in an interpretivist orientation [109]. CDA is particularly useful for unpacking how language constructs and legitimizes particular social realities [62]. It allows for the explicit identification and understanding of discursive practices, particularly by showing how power is exerted and maintained through knowledge construction [39]. It also enables deeper and broader insights into the "why" and "how" when interrogating systems that reinforce and obscure power relations, in ways that other methods may not fully engage with [17]. Finally, CDA helps establish how certain practices and identities become normalized or dominant, often without necessary critical examination [39]. In this case, it allowed us to analyze how companies developing AI tools frame safety in their public-facing materials. Below, we describe our data collection and analysis in more detail.

4.1 Data Collection

For selecting companies, we referred to market share data from StatCounter[1] and Statista[2]. Based on this information, we selected three prominent companies with widely used GenAI based chatbots: OpenAI (ChatGPT), Google (Gemini), and Anthropic (Claude).[3]

Our initial focus was on materials that explicitly mentioned "safety," but broader reading and team deliberation revealed that related concerns—such as trust, responsibility, and harm, were pervasive in the discourse. Scholars have similarly argued for a broader definition of safety that goes beyond the mere absence of harm [126]. Following this perspective, safety can be understood as entangled with values such as inclusion, equitable access, transparency, and accountability, and as socially and politically situated, shaped by structural and systemic conditions, including platform practices and policies [8]. This understanding guided our decision to include materials beyond explicit mentions of "safety," since AI safety is multidimensional, encompassing not only formal policies but also related practices, governance decisions, and indirect mechanisms through which companies manage risk and influence outcomes.

Building on prior work on safety discourse [47], we focused on seven keywords: harm, responsibility, trust, safety, transparency, accountability, and mitigation. Using Google Advanced Search in incognito mode, we applied the query: *site:<company domain> harm OR trust OR safety OR transparency OR accountability OR mitigation OR responsibility*. While not exhaustive, these keywords generally captured how AI platforms frame safety: "safety" and "harm" refer to preventing abuse or misuse of AI, "trust" reflects user confidence in AI behavior, "responsibility" and "accountability" highlight the platform's obligations to manage risks, "transparency" signals how policies and interventions are communicated, and "mitigation" points to concrete strategies such as moderation, verification, or model adjustments. OpenAI's Transparency Report illustrates an example of this framing, explicitly linking evolving policies, monitoring of misuse, and intervention strategies to safety:

> OpenAI's Transparency Report (the "Report"), provided in accordance with the EU Digital Services Act ("DSA"). The Report includes data for content, users and reporters, as applicable, from EU member states, covering the period from February 17, 2024 through

[1] https://gs.statcounter.com/ai-chatbot-market-share
[2] https://www.statista.com/statistics/1618020/ai-chatbots-traffic-share-ww/
[3] Although we initially considered Microsoft, our early analyses revealed that most of their documents focused on business-to-business (B2B) communications, so we decided not to include them. We also did not select Perplexity, as it is a generative search engine rather than a conversational agent unlike the others.

> December 31, 2024 (the "reporting period"). Our policies and practices continue to evolve in conjunction with our services themselves, as well as environmental factors and patterns of potential abuse. We welcome the opportunity to discuss such changes through future reports." (10)

For each company, we collected the top 25 search results. These documents were manually reviewed and filtered for relevance. We included all publicly available documents from company domains that addressed (with or without explicit mention) at least one of the safety-related keywords, without limiting by document type or model— which meant that these included opinions about broader AI safety, or ideas about their own AI models and platforms. We excluded duplicates, broken pages, content irrelevant to safety (e.g., marketing or unrelated technical updates), and externally authored material, ensuring the corpus reflected official, institutionally endorsed discourse on AI safety.

In total, our corpus included 38 *Updates* (blog-style posts on policy changes, committee formations, other safety measures), 9 *Help & Research* documents (guidance and research-oriented materials), and 28 *Perspective* pieces (interpretive or opinion-driven texts shaping broader narrative and discourse). While we cannot state with certainty the intended audiences for these materials, they are official documents affiliated with the respective companies, meaning that users, developers, journalists, and policymakers could regard them as authoritative. For pages with dynamic content, additional screenshots were captured. All data collection was completed by July 2025. The final set of documents was compiled in a spreadsheet and imported into Atlas.ti for qualitative coding and analysis.[4]

We focused on company-authored documents from their official websites because they represent the most stable, public, and institutionally endorsed articulations of corporate positions [12]. Using the company domain address in our search ensured that all captured documents originated from official company domains, maintaining both inclusivity of relevant materials and analytic consistency. This approach avoided including individually authored or external commentary that may not reflect the company's institutional stance, while allowing us to systematically analyze curated outputs such as blog posts, research updates, press releases, and policy statements. Following prior research that similarly analyzed policy and governance documents to trace corporate discourse and value articulation [24, 47, 91], we limited our corpus to these materials to ensure analytic consistency, and traceability of claims within official communication channels.

4.2 Data Analysis

We followed Jäger and Maier [62] in conducting CDA. Our approach combined structural, detailed, and synoptic analyses to examine how companies framed safety in their public-facing documents.[5]

First, we conducted an iterative structural analysis to examine how the texts were organized, including the conceptualization of safety, particularly in terms of the content discussed when companies claimed to address safety. Selected examples of structural codes include *Consequence of Unsafe Usage*, *Developer Responsibility*, *Enforcement Limitations*, and *Mitigation Measures*. The first author conducted an initial reading of a randomly selected subset of safety-related documents and developed a preliminary set of structural codes. These codes were then discussed and refined with the second author, producing a revised codebook. Both authors independently coded five additional randomly selected documents per platform, after which the codebook was further refined for clarity and consistency; this process was repeated twice. The first author then used the finalized codebook to code the full dataset using Atlas.ti, ensuring a systematic and collaborative analysis of all documents.

Building on the structural analysis, the detailed analysis focused on language and discursive strategies. During initial coding, additional codes were generated to capture recurring patterns of discursive positions and framing strategies beyond structure or content, capturing how platforms established and legitimized safety. The first author discussed and refined these codes with the second author. Using this refined set of codes, the first author conducted a comprehensive analysis of the entire dataset. Selected examples include *Cautious Innovation*, *Metaphors*, and *Dynamicity*.

Finally, a synoptic analysis was conducted to integrate and compare insights from earlier stages of the study [62]. For this process, we reviewed the structural codes to understand the overall organization of safety-related discourse across the company-authored documents. Then, we examined the detailed codes to identify specific themes, patterns, and nuances in how safety, trust, responsibility, harm, transparency, accountability, and mitigation were articulated. During this process, we compared patterns across codes and documents, paying particular attention to both explicit statements (what was said) and implicit meanings (how it was said), including the reasons behind explicit statements, possible alternative interpretations, omissions, framing strategies, and the positioning of actors within the discourse. Finally, the insights from this review were synthesized collaboratively among the authors to reconstruct the broader assumptions and knowledge embedded in the corporate discourse, highlighting how companies conceptualized and communicated safety within broader institutional, social, and organizational contexts, and implications for HCI and other fields.

5 Findings

We organize our findings in two separate—yet interrelated—parts (see Table 1). The first part draws on interpretive analyses of our structural codes and examines how GenAI companies structure safety in their discourse (§5.1), focusing on *what* they discuss when addressing safety. The second part draws on analyses of our detailed codes and explores *how* GenAI platforms discuss safety, highlighting the strategies they employ to shape and control the AI safety discourse (§5.2).

5.1 *What* Do Companies Talk About When They Talk About Safety?

In tracing what companies mean when they invoke "safety," our analysis identified three major dimensions: 1) responsibility and accountability, where companies define their own roles while distributing obligations across others; 2) governance, oversight, and control, where safety is tied to internal structures and external

[4]The full dataset can be found here: https://tinyurl.com/y6dp5n56 All primary source materials are stored in the OSF project's dataset/ folder and consist of original corporate safety documents, preserved in their original format.
[5]The codebook can be found here: https://tinyurl.com/y6dp5n56

Table 1: Summary of Findings on GenAI Companies' AI Safety Discourse

Finding Dimension	Brief Overview
***What* Companies Talk About When They Talk About Safety**	
Responsibility and Accountability	AI safety is framed as a shared responsibility across companies, users, governments, and civil society, emphasizing proactive commitments and technical safeguards while leaving accountability for concrete harms weakly specified.
Governance, Oversight, and Control	Safety is tied to internal governance structures and selective regulatory collaboration, with companies advocating for "surgical" oversight that minimizes external constraints while preserving innovation.
Risk, Uncertainty, and Harm Mitigation	Companies enumerate a broad range of risks—from bias and misuse to catastrophic and existential threats—and emphasize continuous evaluation, red-teaming, monitoring, and iterative mitigation practices.
***How* Companies Talk About AI Safety**	
Constructing Authority	Companies establish legitimacy through claims of technical expertise, ethical positioning, and expert-led governance structures, presenting themselves as indispensable stewards of AI safety.
Dynamic and Global Framing	AI safety is framed as an evolving, iterative process that cannot be fully resolved pre-deployment and as a global concern requiring multi-stakeholder coordination.
Metaphors and Analogies	Safety is articulated through metaphors drawn from high-risk domains (e.g., nuclear power, aviation, CBRN threats) and operational analogies that normalize uncertainty while legitimizing ongoing oversight.

collaborations; and 3) risk, uncertainty, and harm mitigation, where companies acknowledged unpredictability while promoting technical safeguards.

5.1.1 Framing Responsibility and Distributing Accountability. Our analysis of accountability and responsibility traces three interrelated dimensions: (1) companies themselves defining and enacting responsibility through technical safeguards and corporate policies; (2) the shifting of responsibility to users and other stakeholders, including governments; and (3) the implications of distributed responsibility for accountability. Together, these dimensions illustrate how companies projected an image of proactive governance while diffusing accountability for the interdependent AI ecosystem.

Across documents, we saw frequent emphasis on responsibility as central to AI safety, framing it as corporate commitments, technical safeguards, and evaluative processes. OpenAI's communication began with a high-level commitment: *"OpenAI is committed to keeping powerful AI safe and broadly beneficial"* (15), going as far as to dictate *Our primary fiduciary duty is to humanity* (121).

Using a similar strategy, Anthropic struck a more process-oriented tone: *"We're committed to evolving our approach alongside these developments, including adapting our frameworks, refining our assessment methods, and learning from both successes and failures along the way."* (29)

OpenAI further reassured enterprise customers that *"we don't train our models on your organization's data by default"* (4), positioning responsibility as a commitment to data-handling practices. They further noted that *"like any technology, these tools come with real risks—so we work to ensure safety is built into our system"* (15), highlighting proactive responsibility-taking. This proactive—and rather technical—responsibility-taking process was also featured in OpenAI's commitment to test their models and conduct red-teaming "before release" (19) aligning with broader shifts in which tech companies moved from the denial of responsibility toward articulating it through ethical commitments[101].

At the same time, companies frequently emphasized that responsibility should be distributed across multiple actors, often involving internal governance structures, authority figures, and external stakeholders. Anthropic described *"independent checks"* and *"assurance structures"* modeled after high-risk industries (43, 45), emphasizing that *"We know we can't do this work alone. We invite researchers, policy experts, and industry partners to collaborate with us as we continue exploring these important questions"* (29). OpenAI relied on committees, preparedness frameworks, and expert advisors (3, 19). Google highlighted the need for *"concrete, context-specific guidance from governments and civil society"* (63), but noted that *"ultimately it is companies and developers who are at the frontline of defense from bad actors"* (63).

Responsibility also extended to users, who were expected to adapt to the perceived increasingly capable AI. Companies discussed the importance of caring for *"psychological factors"* (63), such as automation bias and algorithm aversion, to ensure users do not *"place more faith in its correctness than is warranted"* (63) or *"ignore safety-critical guidance"* (63). Policies, such as Google's Generative AI Prohibited Use Policy (68), explicitly outline how users should engage with AI responsibly, transferring part of the safety burden to end-users. Reporting mechanisms further encouraged user participation in mitigating AI risks (21).

Governments and other public actors were often explicitly mentioned when discussing responsibility, particularly in the context of AI deployment within society. OpenAI stressed collaboration between *"policymakers and AI providers [to] ensure that AI development and deployment is governed effectively at a global scale"* (15), and Anthropic urged governments to act promptly on AI policy (47). Similarly, Google acknowledged that governments and civil society ultimately shape the frameworks within which AI systems are deployed (63). These calls positioned public actors as essential to maintaining safe and responsible AI.

On the other hand, accountability, defined as the obligation to explain and justify the actual harms caused by GenAI models [16], remained ambiguously defined. That is in contrast to forward-looking responsibility, which companies portrayed as a shared duty among

firms, regulators, users, and civil society. In that context, Google described,

> *"Lead on and help shape responsible governance, accountability, and regulation that encourages innovation and maximizes the benefits of AI while mitigating risks (e.g., our role in setting up Partnership on AI, our support for Global Partnership on Artificial Intelligence and our contributions to flagship AI governance efforts, including the EU AI Act, NIST AI Risk Management Framework, and OECD AI Principles)."* (58)

While this emphasized proactive measures, governance, and risk mitigation, it did not clarify who would actually face consequences when harms occur. By framing responsibility separately from enforceable accountability, companies presented themselves as diligent while leaving the question of real-world consequences largely unresolved.

5.1.2 Governance, Oversight, and Control. The discourse around AI safety was also structured through the promotion of internal and external governance structures, which combined public-facing policies with private coordination. By calling for partnerships and participatory approaches to safety, AI companies signaled openness in recognizing that *"getting AI right requires a collective effort"* (58). Furthermore, companies showed support for independent assessments as a way to address limitations in their own AI evaluations.

Companies developed its own internal governance structures for AI safety. These structures took the form of dedicated teams, such as OpenAI's Safety and Security Committee (3), and formal policies for managing risk, like Anthropic's *"clearly-articulated policy on catastrophic risks" (43).* Companies often defined specific public-facing criteria and multi-stage testing processes to evaluate whether a particular AI model was safe, as exemplified by Anthropic's *"Red Line Capabilities"* framework and AI Safety Levels (43).

Although companies promoted their own standards, they positioned themselves as collaborators with governments and policymakers. They framed enforceable regulations, including public transparency requirements for risk policies and evaluation results, as necessary to build trust and ensure accountability. This discourse emphasized the need for a collective, multi-stakeholder approach to effective regulation.

Importantly, the documents also emphasized that regulations should be *"surgical"* and *"simple to understand"* (47) to avoid stifling innovation. Companies were openly cautious about how much external stakeholders could mediate, stating:

> *"Overall, Google recommends a cautious approach for governments regarding liability for AI systems, since the wrong frameworks might place unfair blame, stifle innovation, or even reduce safety. Any changes to the general liability framework should come only after thorough research establishing the failure of the existing contract, tort, and other laws."* (63)

Google employed notably assertive language, explicitly recommending a pro-innovation approach in commentary on the U.S. AI Action Plan (126).Likewise, Anthropic, communicated a similar idea:

> *"Whatever regulations we arrive at should be as surgical as possible. They must not impose burdens that are unnecessary or unrelated to the issues at hand. One of the worst things that could happen to the cause of catastrophic risk prevention is a link forming between regulation that's needed to prevent risks and burdensome or illogical rules."* (47)

This call for some kind of external governance structure also included independent third-party evaluations, where companies advocated to inform *"new standards and laws"* (19), or "*to evaluate the effectiveness of our [their] security controls*" (48), signaling lightly that external oversight is essential to managing AI safety risks.

In summary, the documents showed how their internal teams and safety frameworks were core to their business functions. Likewise, we also noted comapnies' willingness to collaborate with governments and policymakers to shape regulations in pro-innovation ways that were tailored to their specific concerns.

5.1.3 Managing Risks and Mitigation Practices. AI safety was also operationalized through processes and evaluation metrics that identified risks and imposed constraints. Companies framed their efforts as part of a continuous cycle of risk management: *identifying and anticipating harms, testing models at defined stages, limiting or steering their use, and iterating on these steps over time.* Within this cycle, transparency, data sharing, and user feedback were positioned as feedback mechanisms that reinforced safety. In effect, *"safety"* was constructed as an ongoing process of proactive identification and mitigation of risks.

Companies frequently identified a wide spectrum of AI-related risks, including complex and potentially catastrophic dangers beyond familiar concerns such as toxicity and bias, often implying which harms to prioritize. Anthropic, for instance, highlighted that AI behaviors may diverge from designer intentions, including *"sycophancy and a stated desire for power"* (35). Google cataloged risks ranging from technical failures to societal harms such as the amplification of biases and the creation of *"information hazards"* through misinformation and nonfactuality (58). These acknowledgments further encompassed existential threats, including *"large-scale devastation through deliberate misuse"* by *"terrorists or state actors to create bioweapons"* or autonomous actions *"contrary to the intent of their designers"* (44), with Google emphasizing the risk of deceptive alignment, in which systems recognize conflicts with human instructions and deliberately circumvent safeguards (116).

In response to these (and also other) concerns, companies emphasized systematic evaluation and testing as central to their mitigation strategies. Anthropic explicitly argued that *"any industry where there are potential harms needs evaluations,"* drawing analogies to nuclear power monitoring and aircraft flight testing (45). Methods included red-teaming to stress-test models and creating defenses against adversarial misuse. Anthropic advertised its *"Constitutional Classifiers"*, which were designed to guard against *"universal jailbreaks,"* i.e., attempts to elicit forbidden responses (34). Complementing these defenses, companies deployed monitoring systems to detect subtle patterns of misuse. Anthropic's Clio system, for example, enabled *"privacy-preserving analysis of real-world language model use"* to uncover *"coordinated, sophisticated misuse"* that would be invisible at the level of individual interactions (33).

The discourse of risk mitigation was reinforced through organizational and regulatory commitments, as discussed above. OpenAI formed a dedicated safety committee *"responsible for making recommendations on critical safety and security decisions"* (3), while Anthropic created a *"Responsible Scaling Team"* (43). Google emphasized enforcement through its *"Generative AI Prohibited Use Policy"* (68), and all companies pointed to compliance with governmental regulatory frameworks, such as the European Union's GDPR and the US HIPAA (4). Transparency reports detailing banned accounts, government requests, and key metrics were presented as evidence of responsibility and industry leadership.

The last step of the *"safety cycle"* refers to the role of iterative feedback in promoting AI safety. Google emphasized that they *"listen, learn and improve based on feedback from developers, users, experts, governments, and representatives of affected communities [..] and involve human raters to evaluate AI models"* (58), noting that *"human users provide essential feedback to improve AI systems over time"* (63).

This process of feedback closed the loop in the safety cycle: risks were first acknowledged, investigated through evaluation and testing, countered with mitigation tools and governance structures, and continuously refined through user feedback. This iterative process allowed companies to present continuous model refinement and adaptation of safety measures, while signaling responsiveness to real-world use and evolving risks.

5.2 *How* Do Companies Talk About AI Safety?

5.2.1 Constructing Authoritative AI Safety. Weber [141, 142] defined authority as legitimate power accepted by others, based on tradition, charisma, or legal-rational rules. In our corpus, we observe such power as discursively constructed. Authority was enacted through narratives, institutional arrangements, and public statements that demonstrated how these companies were uniquely positioned to manage both the technical and societal consequences of AI, performing legitimacy through strategic narratives and positioning—such as OpenAI talking about its own charter: *"The timeline to AGI remains uncertain, but our Charter will guide us in acting in the best interests of humanity throughout its development."* (121)

Companies constructed authority through technical expertise, ethical positioning, and formal governance. Boards, committees, and expert networks reinforced legitimacy, while forward-looking practices like risk assessment signaled the capacity to anticipate harms. Together, these elements positioned companies as essential stewards of AI, combining expertise, ethics, and institutional oversight.

OpenAI constructed authority by framing itself as the steward of managing societal risks from advanced AI. The company stated that *"to be effective at addressing AGI's impact on society, OpenAI must be on the cutting edge of AI capabilities—policy and safety advocacy alone would be insufficient"* (121), signaling that technical expertise and proximity to the technology were prerequisites for legitimate guidance. Authority was further enacted institutionally through the formation of the Safety and Security Committee, led by Board members Bret Taylor (Chair), Adam D'Angelo, Nicole Seligman, and Sam Altman (CEO), and supported by technical and policy experts including Aleksander Madry, Lilian Weng, John Schulman, Matt Knight, and Jakub Pachocki. The committee also consulted former cybersecurity and national security officials, including retired U.S. Army General Paul M. Nakasone, Rob Joyce, and John Carlin (3). By naming prominent experts and invoking its Charter, OpenAI performed epistemic and moral authority while positioning itself as uniquely capable of anticipating and managing AI-related risks. They further reinforced authority through long-term governance narratives, stating that *"The timeline to AGI remains uncertain, but our Charter will guide us in acting in the best interests of humanity throughout its development,"* performing authority through anticipatory governance and ethical framing.

Anthropic discursively constructed authority by emphasizing deliberative governance and expert-guided decision-making. As a Public Benefit Corporation, it framed its mission around *"ensuring a safe transition through transformative AI"* (40). The document claimed:

> *"Anthropic's Long-Term Benefit Trust today announced the appointment of Richard Fontaine, CEO of the Center for a New American Security, as a new member of the Trust. The Long-Term Benefit Trust (LTBT) is an independent body designed to help Anthropic achieve its public benefit mission."*

Authority was further enacted through the LTBT's guiding role in leadership decisions, as highlighted when a trustee noted, *"My conversations with Dario, Daniela, and the LTBT team made clear that Anthropic takes these challenges seriously. As the stakes get higher, the LTBT serves as a valuable mechanism to help Anthropic's leadership navigate critical decisions. I am pleased to lend my expertise to this important work"* (40).

Google discursively constructed authority by framing itself as the central actor in AI governance while circumscribing the role of external oversight. The company argued that *"self- and co-regulatory approaches remained the most effective practical way to address and prevent AI-related problems in the vast majority of instances"* (63), signaling legitimacy by framing its authority as inherently pragmatic and universally applicable, signaling control and expertise while leaving operational specifics ambiguous. Google simultaneously acknowledged extraordinary societal risks, stating that

> *"Some contentious uses of AI could have such a transformational effect on society that relying on companies alone to set standards was inappropriate—not because companies can't be trusted [..] but because to delegate such decisions to companies would be undemocratic."* (63)

By framing these cases as exceptional, Google reinforced its authority in routine governance, presenting itself as the default steward of AI while exercising strategic control over which issues required external oversight.

5.2.2 Dynamic and Global: Multi-Stakeholder Calls for AI Safety. Our analyses showed that companies discursively constructed AI safety as both *dynamic* and *global*. These framings appeared mutually reinforcing: the inherent uncertainty of *dynamic* safety provided the rationale for treating it as *globally* significant, while the worldwide reach and impact of AI reinforced the necessity for

safety to remain adaptive, iterative, and responsive to emergent circumstances.

Safety was framed as dynamic in multiple ways, encompassing both technical and organizational dimensions. OpenAI, for instance, reported that *"after our latest model, GPT-4, finished training, we spent more than 6 months working across the organization to make it safer and more aligned prior to releasing it publicly"* (15). This description emphasized that safety work extended well beyond model training and was deeply embedded in organizational practice, requiring sustained cross-team coordination and evaluation. At the same time, the company recognized the limits of predeployment efforts, acknowledging that,

> *"We work hard to prevent foreseeable risks before deployment, however, there is a limit to what we can learn in a lab. Despite extensive research and testing, we cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it."* (15)

This statement positioned safety as an inherently emergent and contingent phenomenon: because AI systems interact with complex human and societal contexts in unpredictable ways, companies justified ongoing iterative practices as essential for managing uncertainty.

The iterative dimension of safety extended beyond technical refinement to include stakeholder engagement and gradual deployment strategies. OpenAI described its approach as *"cautiously and gradually releas[ing] new AI systems with substantial safeguards in place to a steadily broadening group of people and make continuous improvements based on the lessons we learn"* (15). By releasing systems incrementally and learning from user interactions, the company framed iteration as a means to both mitigate unforeseen harms and generate empirical knowledge of system behavior in practice. Iteration was also explicitly linked to public participation:

> *"Society must have time to update and adjust to increasingly capable AI, and [...] everyone who is affected by this technology should have a significant say in how AI develops further. Iterative deployment has helped us bring various stakeholders into the conversation about the adoption of AI technology more effectively than if they hadn't had firsthand experience with these tools."* (15)

Other companies echoed this adaptive framing, further emphasizing the evolving nature of AI safety. Anthropic described its approach as *"still evolving. We're sharing our current thinking while acknowledging it will continue to develop as we learn more. We welcome collaboration from across the AI ecosystem as we work to make these systems benefit humanity"* (29). Anthropic also emphasized the formalization of this iterative process through its Risk and Safety Plan, noting that *"We regularly measure the capabilities of our models and rethink our security and safety approaches in light of how things have developed"* (47).

Google similarly highlighted the evolving nature of AI risk and the need for continuous monitoring, stating that *"AI comes with complexities and risks, and these will change over time. As an early-stage technology, its evolving capabilities and uses create potential for misapplication, misuse, and unintended or unforeseen consequences"* (58). To manage these dynamically occurring risks, Google claimed to *"develop methods to monitor deployed systems, ensuring that we can quickly mitigate dynamically-occurring risks in production and in-use services"* (58). Across these examples, safety was constructed as a disciplined—yet flexible—practice, described as *"principled and adaptable to keep up with the evolving AI landscape"* (29). Collectively, these statements framed safety as an ongoing process of learning, adaptation, and reflection rather than as a static goal that could be fully achieved at a single point in time.

The *dynamic* framing of safety provided the rationale for presenting it as *global*. Because AI systems could produce unintended consequences that extended beyond localized contexts, companies emphasized the worldwide stakes of safety management. Google's discourse explicated:

> *"Our approach to developing harnessing the potential of AI is grounded in our founding mission to organize the world's information and make it universally accessible and useful and it is shaped by our commitment to improve the lives of as many people as possible. It is our view that AI is now, and more than ever, critical to delivering on that mission and commitment."* (58)

OpenAI, on the other hand, underscored the scope of its responsibility, noting that *"more than a hundred million users and millions of developers rely on the work of our safety teams"* (19). By emphasizing the large, dispersed population affected by AI systems, the company framed safety as consequential not only for individual users but also for broader societal and technological ecosystems. Google extended this reasoning, describing responsible AI as *"a collective effort [..] involving researchers, developers, users (individuals, businesses, and other organizations), governments, regulators, and citizens"* (58).

5.2.3 Metaphors. Our analyses revealed that companies discursively constructed AI safety by drawing on metaphors, historical analogies, and procedural frameworks to convey its stakes. A metaphor is a way of understanding one thing in terms of another [76], and in this context, metaphors can reveal specific comparisons that shape baseline ways of understanding what safe AI is and how it can be operationalized [63]. While most metaphors were not explicated in boundless ways, they implied comparisons to other technologies that created particular imaginaries.

First, safety was consistently constructed through comparisons to dual-use technologies and catastrophic scenarios. Anthropic, for example, highlighted the potential of AI to amplify threats in domains traditionally associated with existential risk, positioning it also as generating extreme and unpredictable harms:

> *"Chemical, biological, radiological and nuclear (CBRN) risks—We're prioritizing evaluations that assess two critical capabilities: a) the potential for models to significantly enhance the abilities of non-experts or experts in creating CBRN threats, and b) the capacity to design novel, more harmful CBRN threats."* (27)

OpenAI likewise underscored the temporal and strategic dimensions of oversight, framing safety as a long-term investment: *"We view safety as something we have to invest in and succeed at across multiple time horizons, from aligning today's models to the far more capable systems we expect in the future"* (19). OpenAI also described

its release of ChatGPT as a *"Rorschach test,"* a metaphor suggesting that safety itself was interpreted through projection: depending on whether one believed AI progress to be continuous or discontinuous, the release appeared either as a reckless gamble or as an opportunity to learn (9). Collectively, these framings constructed a landscape dominated by hypothetical threats, foregrounding the scale and unpredictability of potential harms while leaving the question of practical mitigation largely unaddressed.

The metaphorization of AI safety was reinforced through technological and historical analogies. Anthropic compared AI safety practices to established safety regimes in other high-risk industries: *"Nuclear power stations have continuous radiation monitoring and regular site inspections; new aircraft undergo extensive flight tests to prove their airworthiness"* (45). By linking AI to domains with robust regulatory and inspection frameworks, companies constructed a discourse in which AI was both societally consequential and technically tractable but only under sustained, rigorous oversight. Google elaborated on this approach through references to transformative technologies of the past: *"In thinking through these issues, it may be helpful to review how the world has responded to the emergence of other technologies presenting ethical (and at the extreme, existential) questions"* (63). They cited genetic engineering, in vitro fertilization, nuclear technology, PCBs, and space exploration as analogues, demonstrating how voluntary and multilateral norms had historically mitigated risks while enabling innovation (63). For instance, the Asilomar Conference on Recombinant DNA and the Outer Space Treaty were invoked to exemplify how collective international coordination could provide frameworks for responsible governance (63). These analogies also functioned rhetorically, framing companies as inheritors and interpreters of past governance successes and failures, tasked with applying these lessons to a novel, high-stakes technology.

Operational metaphors further concretized AI's risks and the challenges of alignment. Google's examples of misaligned agents illustrated both the technical and ethical dimensions of safety: *"Suppose a cleaning robot maker set the objective to remove visible dirt as fast as possible. If the optimal approach turned out to be hiding dirt under the carpet, or throwing away all visible dirty objects, this would be a failure in spirit even though it might satisfy the objective"* (63). A similar scenario, *"a robot barista tasked with delivering coffee in the shortest time possible might (if given free rein) come up with the solution to throw the cup!"* (63), highlighted the tension between optimizing objective functions and unintended harmful outcomes. These examples translated abstract alignment problems into everyday scenarios, legitimizing the need for safeguards, constrained exploration, and iterative testing. Google further connected these operational metaphors to practices for establishing standards of explanation and due diligence: *"Researchers from the public, private, and academic sectors should work together to outline basic workflows and standards of documentation for specific application contexts which would be sufficient to show due diligence in carrying out safety checks (e.g., like for airline maintenance)"* (63). In other words, operational metaphors provided a bridge between theoretical risk and the practical procedures companies were using to manage it.

6 Discussion

Our findings illuminate both what was discussed within safety and safety-adjacent documents and how these discussions were framed. We found that responsibility, accountability, governance, and risk mitigation constituted core thematic concerns in articulations of AI safety. Furthermore, we observed an implicit form of authority construction in which corporate or corporate-state entities were positioned as the appropriate stewards of AI safety and regulation. Although corporate discourse on AI safety included outside actors, such as governments, civil society, and practitioners, the power to set priorities and the epistemic authority remained firmly within the companies. These documents also created leeway for failures by emphasizing the inherent unpredictability of AI systems. Finally, they relied heavily on technological analogies as metaphors to frame AI as both vast and difficult to control. Within HCI, we see our work as formative and exploratory, with potential contributions to AI literacy and governance in particular. We contextualize our findings in the following sections.

6.1 AI (Safety) Literacy

Efforts around AI literacy within HCI have largely centered on helping users understand what AI is and how AI systems function [81]. Existing work has emphasized explainability [11, 28, 66] and mental models [22], aiming to make machine learning systems more legible to non-expert users through transparency mechanisms, interfaces, and pedagogical tools [69, 70]. Parallel work in explainable AI (XAI) has sought to provide post-hoc rationales for algorithmic outputs [33], supporting trust calibration, debugging, and user comprehension. At the same time, in tangential academic communities, substantial effort has been dedicated to auditing [26, 48], measuring, and mitigating AI harms through benchmarks, red-teaming, and evaluation frameworks [7, 87]. Yet, users develop their understanding of AI and their risks not only through direct use, but also through company communications, media, and educational materials [83], including those about AI safety—which we have systematically studied.

Our work extends prevailing notions of AI literacy by arguing that it should move beyond technical understanding and output evaluation to encompass a critical, discursive dimension, enabling individuals to examine AI's ethical implications, societal impact, underlying assumptions, and potential alternatives within the sociotechnological system [136]. As shown in our analysis, AI companies seek to explicate their AI safety practices, position themselves as authorities, and set priorities for future AI development. Being public-facing, these documents often function as primary sites through which companies communicate their safety commitments to users, making them powerful pedagogical artifacts in themselves. Reading them critically is therefore foundational to AI safety literacy.

Importantly, this critical-discursive framing aligns with policy efforts. The European Union's AI Act explicitly defines AI literacy as the capacity to understand potential risks and harms associated with AI [36] and calls for promoting AI literacy and public awareness through multiple channels, including training, guidelines, and best practices for stakeholders interacting with AI systems [37]. By complementing these regulatory objectives, a discursive approach to AI literacy emphasizes not only comprehension of risks and

opportunities but also the critical evaluation of corporate narratives, rhetorical strategies, and accountability structures embedded in AI governance.

Furthermore, scholars have argued that AI literacy should be framed as a set of competencies [4, 81] such as verifying information against trusted sources, understanding outputs, and documenting or flagging problematic system behavior. Prior HCI work has demonstrated the potential of such practices: Deng et al., [27] engaged stakeholders in end-user contestation of AI claims, while auditing research has examined how LLM outputs can be evaluated within educational contexts [110]. We build on these contributions by explicitly foregrounding the need to interrogate assumptions, omissions, power relations, and political motivations embedded in AI safety narratives.

We introduce the notion of a *discursive toolkit* as a mechanism and artifact to support AI safety literacy. Prior work has shown that toolkits are artifacts that shape organizational imaginaries. For example, Wong et al. demonstrated that AI-ethics toolkits encoded specific institutional visions and influenced how practitioners interpreted responsibility [147]; Hollanek highlighted that diverse toolkits often served symbolic ethico-political purposes limiting their capacity for change [58]; and Krafft et al. developed an AI-policy toolkit for community advocates to open up AI policy as a public problem [74]. However, these toolkits largely focus on designing and deploying AI tools rather than explicitly *communicating* technology. We argue for toolkits that help users identify recurring metaphors, recognize patterns of responsibility diffusion, and question who is positioned as accountable for, for instance, mapping environmental and labor costs associated with prompts [75] or reflecting on how GenAI tools shape decision-making and ethical responsibility [71]. By situating these practices in HCI scholarship on power and accountability [49, 82], AI literacy can move beyond interpreting and knowing the outputs to asking questions about the systems in use, its parent companies, its background workings and critically interrogating the narratives they produce.

Our analysis further shows that AI companies frame safety along multiple, interrelated dimensions, distributing responsibility across internal teams, external experts, users, and governments while leaving accountability for actual harms ambiguously defined. Table 2 summarizes some of these dimensions with concrete examples, linking each to key literacy insights. Because our work is interpretive, the table serves as a *non-exhaustive*, analytic component that could inform the development of future AI literacy toolkits. It illustrates how users could be supported in identifying accountability structures, analyzing corporate safety narratives, and reflecting on the social and ethical implications of AI. By grounding these analytic components in broader AI literacy goals, future toolkits could operationalize interpretive and critical competencies, supporting users in navigating AI safety as a dynamic, socially negotiated, and context-dependent process.

At the same time, caution is needed: GenAI companies are increasingly promoting AI literacy on their own terms,[6] embedding products directly into classrooms and curricula [103]. A critical form of AI literacy must therefore deliberately maintain analytical distance from corporate framings, equipping users to assess claims and institutional agendas rather than accepting them at face value.

6.2 Governance in AI: Discursive Framing, Participatory Tensions, and Corporate Power

AI governance has become a growing focus within HCI, as scholars examine how participatory design, civic engagement, and sociotechnical interventions can support more inclusive and accountable forms of governance. The CHI 2025 Sociotechnical AI Governance workshop, organized by Feng et al. [41], exemplifies this interest. The workshop brought together policymakers, designers, technologists, and affected communities to explore participatory governance models, documentation practices, stakeholder-centered design, and cross-sector communication mechanisms. Papers featured in the workshop highlighted approaches for bridging stakeholder perspectives [148], designing tools to enable literacy for regulators [128], and fostering civic engagement in AI governance [59], all aimed at facilitating meaningful participation and collaborative governance in AI systems. Importantly, these discussions intersect with policy frameworks such as the EU AI Act, which seeks to promote AI literacy, transparency, and risk-based oversight [138].

Building on this scholarly context, our findings show that the same language of participation, collaboration, and multi-stakeholder engagement can be strategically appropriated by AI companies to reinforce their own authority. Companies frequently foreground collaboration in corporate communications. For example, Anthropic states: *"We welcome collaboration from across the AI ecosystem as we work to make these systems benefit humanity"* (29). Yet this rhetoric coexists with resistance to regulatory constraints, exemplified by their caution that regulations should be as surgical as possible, not burdening companies (47).

Such examples illustrate how participatory language can coexist with efforts to limit oversight. While attempts to make AI governance, design, and deployment more participatory exist [25, 132] we argue that corporate-driven communication also functions as a form of governance, shaping public expectations while narrowing the space for accountability. By critically analyzing corporate discourse, our work highlights the risks of co-optation and symbolic participation and challenges the assumption that stakeholder engagement is inherently equitable or meaningful. The ways in which participation is framed shape how people interact with systems, make decisions, and understand their role, meaning that rhetoric itself influences both user experience and the design of participatory processes.

These dynamics resonate with longstanding critiques of self-regulation and corporate influence on policy formation. Our analysis finds that companies frequently portray themselves as self-regulators, emphasizing forward-looking responsibility and proactive management of potential harms—a discursive strategy that signals accountability without binding external oversight. In practice, this often appears through momentary and opportunistic references to *governance,* reflecting what Roberge et al. [117] described as *"governance at a distance: to navigate the complexities of the present, it is deemed better to aim for a horizon that is as remote as possible and hope for the best."* Examples include OpenAI's emphasis

[6] https://academy.openai.com/

Table 2: Exploratory dimensions of AI safety discourse, showing (1) recurrent discursive dimensions identified across company safety documents, (2) representative examples of how each dimension was articulated in practice, and (3) the corresponding AI literacy insights that these articulations afford, revealing how safety is framed, operationalized, and made legible to the public.

Dimension	Examples	AI Literacy Insight
Responsibility & Accountability	Forward-looking commitments, distributed across companies, users, and governments; accountability often vague	Responsibility ≠ enforceable accountability; signals diligence without clear consequences
Governance & Oversight	Internal safety teams, policies, multi-stakeholder collaboration; cautious calls for regulation	Recognize authority construction and regulatory shaping through organizational structures
Risk & Mitigation	Iterative evaluation, testing, red-teaming, monitoring, user feedback	Safety as an ongoing process, not a static property
Dynamic & Global Framing	Safety portrayed as emergent, evolving, and globally consequential	Safety is contingent, context-dependent, and socially negotiated
Metaphors & Analogies	Comparisons to CBRN, nuclear power, aviation; operationalized through everyday scenarios	Connects abstract AI risks to familiar real-world scenarios, highlighting both constraints and potential misunderstandings of AI's impact relative to existing technologies

on multi-horizon safety planning (19) and Anthropic's evolving safety frameworks (29). These discourses illuminate gaps between stated commitments and operational practices, underscoring the importance of mechanisms that translate participatory ideals into actionable oversight because self-regulation and forward-looking responsibilities rarely produce genuine accountability [115].

Moreover, even when formal regulations exist, such as EU AI Act and the revised EU Product Liability Directive, companies can dilute their impact [138] or slow down their implementation [55]. Furthermore, corporate influence extends to shaping policy narratives. EU leaders have repeated corporate talking points, while US regulators echo long-term concerns about AI risks, aligning with industry framing [85]. These examples highlight how companies not only resist regulation but also actively shape regulatory discourse to maintain operational flexibility. Lawsuits over data use and authorship [1, 15] and lobbying around AI regulation demonstrate that voluntary corporate transparency can be incomplete or strategically deployed. Internal governance structures, safety committees, and epistemic authority-building events [3] further illustrate how corporations consolidate control over both AI systems and public narratives about their safety.

Companies also rely on normatively resonant metaphors to communicate risk, safety, and responsibility, constructing and stabilizing organizational culture [44]. GenAI systems are described as "frontier technologies," safety processes likened to "aircraft stress-testing" or "nuclear incident monitoring," and misalignment compared to "containment breaches" or "hazardous-material spills." These metaphors draw on historically familiar narratives of scientific risk, military defense, and crisis management, legitimizing corporate authority while underscoring the stakes involved in AI safety. These linguistic and cultural strategies reinforce the broader patterns of selective engagement with governance, connecting back to the earlier discussion of symbolic participation and corporate framing.

This aligns with Ahmed et al.'s concept of an *epistemic culture* of AI safety [3], in which authority is produced through interconnected networks of institutes, research grants, media, competitions, policy engagement, and career infrastructures. These ecosystems consolidate power within a relatively narrow group of actors, creating an epistemic monoculture in which certain forms of risk, safety, and legitimacy dominate. Sociotechnical imaginaries further illuminate this process [146]: by framing the future of AI as inherently dependent on corporate stewardship, these imaginaries guide both policy and design decisions, embedding assumptions about stakeholders, public benefit, and technological trajectories. Our analysis also resonates with HCI scholarship on *futuring*, showing how speculative visions influence what becomes thinkable, fundable, and legitimate [118]. By repeatedly positioning themselves as guardians of the future, corporate actors crowd out alternative imaginaries, reinforcing the notion that those who build these systems are best positioned to govern them.

HCI scholars must actively challenge these dynamics by rigorously scrutinizing corporate discourse, exposing power asymmetries, and shaping policy debates. Our analyses demonstrate that ostensibly participatory rhetoric often masks symbolic—rather than substantive—engagement, and HCI research is uniquely positioned to make such performative practices visible. Beyond analysis, research communities and conferences have a responsibility to critically examine their own institutional practices, such as sponsorship policies, to safeguard independence and uphold the integrity of participatory AI governance research, following precedents like the FAccT conference's restrictions on corporate sponsorship [6].

For the HCI community, this implies that designing more collaborative or participatory governance processes is insufficient if researchers do not also interrogate the discursive conditions under

which participation is invited, structured, and constrained. Our findings suggest that HCI research on AI governance should more explicitly analyze how corporate rhetoric shapes regulatory imagination, stakeholder roles, and the perceived limits of intervention. By treating discourse itself as a site of governance, HCI scholars can better identify where participatory ideals risk being co-opted—and where alternative sociotechnical arrangements might meaningfully redistribute power.

7 Limitations and Future Work

Our study has several limitations that future work can bridge and extend upon. First, our corpus consists exclusively of publicly available documents and pages produced by the three companies. We did not include internal documentation; as a result, our analysis captures primarily how companies communicate about AI safety externally, rather than the full range of internal practices or informal discussions. While this focus allows us to examine corporate narratives and public-facing discourse, it may not reflect all mechanisms through which safety practices are operationalized. Future work could combine document analysis with interviews, ethnography, or "read-along" studies to better understand how users interpret, navigate, and make sense of corporate safety narratives, as well as include ethnographic work with developers inside these companies. Additionally, the documents were often all-encompassing, speaking interchangeably about both models and platforms. While we retained this in our analysis, future research could more clearly differentiate between them to examine whether particular ideas are communicated through specific artifacts.

Second, we sampled three major platforms, excluding smaller companies, domain-specific systems, and materials beyond our keywords. Broader datasets could reveal additional discursive patterns and framings.

Third, critical discourse analysis is interpretive; despite iterative coding and collaborative checks, our readings remain situated and cannot be generalizable. Complementary methods such as quantitative text analysis, surveys, or controlled experiments could assess how audiences perceive and act on safety discourse. Finally, corporate safety discourse evolves rapidly; longitudinal studies are needed to track changes alongside regulation, market pressures, and public debates.

8 Conclusion

This study offers one of the first empirical examinations of how safety is discursively constructed by corporate actors of generative AI platforms. Using CDA, we analyzed a stratified set of public-facing documents from OpenAI, Google, and Anthropic, focusing on how safety and other similar constructs were discursively constructed. By situating these materials in real-world contexts where users seek information about AI safety, our approach illuminated not only what companies claim about safety, but also how these claims are structured, framed, and legitimized through language.

Our analysis of corporate AI safety discourse reveals how companies construct authority, diffuse accountability, and frame risks, often positioning themselves as the primary stewards of safety while rhetorically invoking collaboration and participation. These findings underscore two key contributions for HCI: first, advancing AI literacy by treating corporate narratives as pedagogical artifacts that should be critically interrogated to understand assumptions, power dynamics, and ethical implications; and second, informing AI governance by showing how participatory language and metaphors can legitimize authority while constraining meaningful oversight or participation.

Acknowledgments

We thank Dr. Kelley Cotter, Dr. Elissa Redmiles, and Dr. Abraham Mhaidli for their invaluable feedback on earlier drafts of this work. We also appreciate the insights and support from our colleagues at the Max Planck Institute for Security and Privacy (MPI-SP), particularly members of the Human-Centered Security and Privacy group, and the LIKED Lab at Penn State. This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 CASA – 390781972.

References

[1] 2023. Getty Images (US), Inc. v. Stability AI, Inc. U.S. District Court for the District of Delaware, Filing 1. Retrieved from https://tinyurl.com/2r7r7a4c.

[2] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. 2024. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294* (2024).

[3] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2024. Field-building and the epistemic culture of AI safety. *First Monday* 29, 4 (2024), –. doi:10.5210/fm.v29i4.13626

[4] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. 2025. Generative AI Literacy: Twelve Defining Competencies. *Digit. Gov.: Res. Pract.* 6, 1, Article 13 (Feb. 2025), 21 pages. doi:10.1145/3685680

[5] Terje Aven and Marja Ylönen. 2018. A risk interpretation of sociotechnical safety perspectives. *Reliability Engineering & System Safety* 177 (2018), 217–226. doi:10.1016/j.ress.2018.03.004

[6] Solon Barocas and Mireille Hildebrandt. 2023. *FAccT Sponsorship Policy Review.* Technical Report. FAccT Executive Committee. Commissioned by FAccT's Executive Committee. As of August 2023, recommendations are pending approval by FAccT's Steering Committee.

[7] Anthony M Barrett, Krystal Jackson, Evan R Murphy, Nada Madkour, and Jessica Newman. 2024. Benchmark early and red team often: A framework for assessing and managing dual-use hazards of ai foundation models. *arXiv preprint arXiv:2405.10986* (2024).

[8] Louisa Bartolo and Ariadna Matamoros-Fernández. 2023. Online Harm. https://law.yale.edu/sites/default/files/area/center/isp/documents/onlineharm_ispessayseries_2023.pdf Accessed: 2025-09-02.

[9] Juana Catalina Becerra Sandoval and Felicia S. Jing. 2025. Rethinking AI Safety: Provocations from the History of Community-based Practices of Road and Driver Safety. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25).* Association for Computing Machinery, New York, NY, USA, 964–974. doi:10.1145/3715275.3732062

[10] Sari Berkovich. 2023. *The History of Trust & Safety.* ActiveFence. https://www.activefence.com/blog/the-history-of-trust-and-safety Accessed: 2025-07-22.

[11] Maalvika Bhat and Duri Long. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24).* Association for Computing Machinery, New York, NY, USA, 939–957. doi:10.1145/3643834.3660722

[12] Kean Birch and Kelly Bronson. 2022. Big Tech. *Science as Culture* 31, 1 (2022), 1–14. doi:10.1080/09505431.2022.2036118

[13] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021).

[14] Susanne Bødker, Myriam Lewkowicz, and Alexander Boden. 2020. What's in a word? Platforms Supporting the Platform Economy. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20).* Association for Computing Machinery, New York, NY, USA, Article 87, 10 pages. doi:10.1145/3419249.3420167

[15] Robert Booth. 2025. Teen killed himself after 'months of encouragement from ChatGPT', lawsuit claims. https://www.theguardian.com/technology/2025/aug/

27/chatgpt-scrutiny-family-teen-killed-himself-sue-open-ai. Accessed: 2025-09-08.

[16] Mark Bovens. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal* 13, 4 (2007), 447–468. doi:10.1111/j.1468-0386.2007.00378.x

[17] André Brock. 2018. Critical technocultural discourse analysis. *New Media & Society* 20, 3 (2018), 1012–1030. doi:10.1177/1461444816677532 arXiv:https://doi.org/10.1177/1461444816677532

[18] Giulia Campaioli, Adriano Zamperini, and Marta Cecchinato. 2025. 'Now you're home': Awareness cues, rejection and post-digital safety on mobile dating apps. *New Media & Society* (2025), 14614448251336437.

[19] Catherine Clifford. 2018. Google CEO: A.I. is more important than fire or electricity. *CNBC Make It* (1 February 2018). https://www.cnbc.com/2018/02/01/google-ceo-ai-is-more-important-than-fire-or-electricity.html Accessed: 2025-09-04.

[20] Rachel Coldicutt. 2024. AI safety is a narrative problem. *Harvard Data Science Review* Special Issue 5 (2024).

[21] Lizzie Coles-Kemp, Rikke Bjerg Jensen, and Claude PR Heath. 2020. Too much information: Questioning security in a post-digital society. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[22] Aayushi Dangol, Robert Wolfe, Runhua Zhao, JaeWon Kim, Trushaa Ramanan, Katie Davis, and Julie A. Kientz. 2025. Children's Mental Models of AI Reasoning: Implications for AI Literacy Education. In *Proceedings of the 24th Interaction Design and Children (IDC '25)*. Association for Computing Machinery, New York, NY, USA, 106–123. doi:10.1145/3713043.3728856

[23] Ankolika De and Kelley Cotter. 2026. The discursive flexibility of changecraft: Platform change discourse in Meta, TikTok, YouTube, and X. *Platforms & Society* 3 (2026), 29768624251408212. doi:10.1177/29768624251408212

[24] Julia R. DeCook, Kelley Cotter, Shaheen Kanthawala, and Kali Foyle. 2022. Safe from "harm": The governance of violence by platforms. *Policy & Internet* 14, 1 (2022), 63–78. doi:10.1002/poi3.290

[25] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)* (Boston, MA, USA). ACM, 23. doi:10.1145/3617694.3623261

[26] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW521 (Oct. 2025), 35 pages. doi:10.1145/3757702

[27] Wesley Hanwen Deng, Michelle S. Lam, Ángel Alexander Cabrera, Danaë Metaxa, Motahhare Eslami, and Kenneth Holstein. 2023. Supporting User Engagement in Testing, Auditing, and Contesting AI. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 556–559. doi:10.1145/3584931.3611279

[28] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1591–1602. doi:10.1145/3461778.3462131

[29] Guy Dishon. 2024. From Monsters to Mazes: Sociotechnical Imaginaries of AI Between Frankenstein and Kafka. *Postdigital Science and Education* 6 (2024), 962–977. doi:10.1007/s42438-024-00482-4

[30] Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating algorithmic bias: From neutrality to justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 691–704.

[31] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–32.

[32] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The algorithmic imprint. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1305–1317.

[33] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 263–274. doi:10.1145/3301275.3302316

[34] Elizabeth Ellcessor. 2022. Maps and the Affective Surveillance of "Safety". In *In Case of Emergency: How Technologies Mediate Crisis and Normalize Inequality*. NYU Press, 44–68. https://www.jstor.org/stable/jj.4493291.6

[35] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. *arXiv preprint arXiv:2502.06559* (2025).

[36] European Union. 2024. Artificial Intelligence Act (Regulation (EU) 2024/1689). Regulation of the European Parliament and of the Council. Official Journal of the European Union. Article 3(56) defines "AI literacy" as the skills, knowledge and understanding enabling informed deployment of AI systems and awareness of associated risks and harms.

[37] European Union. 2024. Artificial Intelligence Act (Regulation (EU) 2024/1689), Article 66: Tasks of the Board. Official Journal of the European Union. The European Artificial Intelligence Board supports the Commission in promoting AI literacy, public awareness, and understanding of AI risks, benefits, safeguards, and stakeholder responsibilities.

[38] Cornelia Evers. 2024. Talking past each other? Navigating discourse on ethical AI: Comparing the discourse on ethical AI policy by Big Tech companies and the European Commission. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1885–1896. doi:10.1145/3630106.3659013

[39] Norman Fairclough. 2013. *Critical Discourse Analysis: The Critical Study of Language*. Taylor & Francis, United Kingdom.

[40] Conor Feehly. 2025. Truth, Romance and the Divine: How AI Chatbots May Fuel Psychotic Thinking. *Scientific American* (24 August 2025). https://www.scientificamerican.com/article/truth-romance-and-the-divine-how-ai-chatbots-may-fuel-psychotic-thinking/ Accessed: 2025-09-06.

[41] KJ Kevin Feng, Rock Yuren Pang, Tzu-Sheng Kuo, Amy Winecoff, Emily Tseng, David Gray Widder, Harini Suresh, Katharina Reinecke, and Amy X Zhang. 2025. Sociotechnical AI Governance: Challenges and Opportunities for HCI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[42] Andrea Ferrario, Alberto Termine, and Alessandro Facchini. 2024. Addressing Social Misattributions of Large Language Models: An HCXAI-based Approach. *arXiv preprint arXiv:2403.17873* (2024). arXiv:2403.17873 [cs.AI] https://doi.org/10.48550/arXiv.2403.17873 Extended version accepted for the ACM CHI Workshop on Human-Centered Explainable AI 2024 (HCXAI24).

[43] Gabriele Ferri and Inte Gloerich. 2023. Risk and harm: Unpacking ideologies in the AI discourse. In *Proceedings of the 5th International conference on Conversational User Interfaces*. 1–6.

[44] S. Förster and Y. Skop. 2025. Between fact and fairy: Tracing the hallucination metaphor in AI discourse. *AI & Society* (2025). doi:10.1007/s00146-025-02392-w

[45] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* 5, 1 (2019), 2056305119832588. doi:10.1177/2056305119832588 arXiv:https://doi.org/10.1177/2056305119832588

[46] Tarleton Gillespie, Ryland Shaw, Mary L Gray, and Jina Suh. 2024. AI red-teaming is a sociotechnical challenge: on values, labor, and harms. *arXiv preprint arXiv:2412.09751* (2024).

[47] Rosalie Gillett, Zahra Stardust, and Jean Burgess. 2022. Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions. *Social Media + Society* 8, 4 (2022), 20563051221144315. doi:10.1177/20563051221144315 arXiv:https://doi.org/10.1177/20563051221144315

[48] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069* (2019). doi:10.48550/arXiv.1806.00069 Presented at the 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018).

[49] Xingjian (Lance) Gu and Barbara J. Ericson. 2025. AI Literacy in K-12 and Higher Education in the Wake of Generative AI: An Integrative Review. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.1 (ICER '25)*. Association for Computing Machinery, New York, NY, USA, 125–140. doi:10.1145/3702652.3744217

[50] Eileen Guo, Geiger Geiger, and Justin-Casimir Braun. 2025. Inside Amsterdam's high-stakes experiment to create fair welfare AI. MIT Technology Review. https://tinyurl.com/44s63vuk.

[51] Thilo Hagendorff. 2024. Mapping the ethics of generative AI: A comprehensive scoping review. *Minds and Machines* 34, 4 (2024), 39.

[52] Jacqueline Harding and Cameron Domenico Kirk-Giannini. 2025. What is AI safety? What do we want it to be? *Philosophical Studies* 182, 7 (July 2025), 1495–1518. doi:10.1007/s11098-025-02367-z Published 24 June 2025; Open access under CC BY 4.0.

[53] Cynthia Hardy and Robyn Thomas. 2015. Discourse in a Material World. *Journal of Management Studies* 52, 5 (2015), 680–696. doi:10.1111/joms.12113 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/joms.12113

[54] Bridget A Harris and Delanie Woodlock. 2019. Digital coercive control: Insights from two landmark domestic violence studies. *The British Journal of Criminology* 59, 3 (2019), 530–550.

[55] Robert Hart and Dominic Preston. 2025. Europe is scaling back its landmark privacy and AI laws. *The Verge* (2025). https://www.theverge.com/news/823750/european-union-ai-act-gdpr-changes

[56] Gina Helfrich. 2024. The harms of terminology: why we should reject so-called "frontier AI". *AI and Ethics* 4, 3 (2024), 699–705.

[57] Mél Hogan. 2024. AI is a hot mess. *Training the archive* (2024), 33–54.

[58] Tomasz Hollanek. 2024. The Ethico-Politics of Design Toolkits: Responsible AI Tools, From Big Tech Guidelines to Feminist Ideation Cards (Extended Abstract). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '24, Vol. 7)*. ACM, 610–610. doi:10.1609/aies.v7i1.31663

[59] Mst Rafia Islam and Azmine Toushik Wasi. 2025. Public Participation in AI Governance: A Meta-Analysis of Citizen Engagement Initiatives. In *STAIG@CHI'25 Accepted Papers*. CHI Conference on Human Factors in Computing Systems. Accepted paper analyzing citizen engagement initiatives in AI governance.

[60] Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, Anthony Tung, Xiaokui Xiao, and Audrey Yue. 2025. Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. *Digit. Gov.: Res. Pract.* 6, 1, Article 11 (Feb. 2025), 15 pages. doi:10.1145/3689372

[61] Mackenzie Jorgensen, Hannah Richert, Elizabeth Black, Natalia Criado, and Jose Such. 2023. Not so fair: The impact of presumably fair machine learning models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 297–311.

[62] Siegfried Jäger and Florentine Maier. 2014. *Analysing discourses and dispositives: A Foucauldian approach to theory and methodology*. https://www.diss-duisburg.de/2014/06/analysing-discourses-and-dispositives/ Noch nicht begutachtetes Draft: Ruth Wodak/Michael Meyer: *Methods of Critical Discourse Analysis*, 3rd ed., Sage Publications Ltd, London 2014.

[63] Kaisla Kajava and Nitin Sawhney. 2023. Language of Algorithms: Agency, Metaphors, and Deliberations in AI Discourses. In *Handbook of Critical Studies of Artificial Intelligence*, Simon Lindgren (Ed.). Edward Elgar, 1–15. doi:10.4337/9781803928562.00025

[64] Cecilia Kang. 2023. OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times* (2023). https://www.nytimes.com/2023/05/16/technology/openai-altman-senate-hearing.html A version appeared in print on May 17, 2023, Section B, Page 1.

[65] Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. Epistemic injustice in generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 684–697.

[66] Patrick Gage Kelley and Allison Woodruff. 2023. Advancing Explainability Through AI Literacy and Design Resources. *Interactions* 30, 5 (Aug. 2023), 34–38. doi:10.1145/3613249

[67] Os Keyes, Josephine Hoy, and Margaret Drouhard. 2019. Human-Computer Insurrection: Notes on an Anarchist HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300569

[68] Shaleen Khanal, Hongzhou Zhang, and Araz Taeihagh. 2025. Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society* 44, 1 (Jan. 2025), 52–69. doi:10.1093/polsoc/puae012

[69] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 822–835.

[70] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[71] Shamika Klassen and Casey Fiesler. 2022. "Run Wild a Little With Your Imagination": Ethical Speculation in Computing Education with Black Mirror. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for Computing Machinery, New York, NY, USA, 836–842. doi:10.1145/3478431.3499308

[72] Youjin Kong. 2022. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 485–494.

[73] Youjin Kong. 2025. What Is the Point of Equality in Machine Learning Fairness? Beyond Equality of Opportunity. *arXiv preprint arXiv:2506.16782* (2025).

[74] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. doi:10.1145/3442188.3445938

[75] Priya C Kumar, Kelley Cotter, and Laura Y Cabrera. 2024. Taking responsibility for meaning and mattering: An agential realist approach to generative AI and literacy. *Reading Research Quarterly* 59, 4 (2024), 570–578.

[76] George Lakoff and Mark Johnson. 2003. *Metaphors We Live By* (updated with afterword ed.). University of Chicago Press, Chicago, IL.

[77] Scott Lash. 2015. Performativity or Discourse? An Interview with John Searle. *Theory, Culture & Society* 32, 3 (2015), 135–147. doi:10.1177/0263276415571940 arXiv:https://doi.org/10.1177/0263276415571940

[78] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? 138–138 pages.

[79] Jinsook Lee, Emma Harvey, Joyce Zhou, Nikhil Garg, Thorsten Joachims, and Rene F Kizilcec. 2024. Algorithms for College Admissions Decision Support: Impacts of Policy Change and Inherent Variability. *arXiv preprint arXiv:2407.11199* (2024).

[80] Janne Martha Lentz, Christiane Meyer-Habighorst, Mê-Linh Riemann, Anke Strüver, Sarah Baumgartner, Sarah Staubli, Nicola Techel, Sybille Bauriedl, and Karin Schwiter. 0. From Exceptionalism to Normalisation: How Narratives of Platform Companies Legitimise Precarious Work and Commodified Care. *Critical Sociology* 0, 0 (0), 08969205241306300. doi:10.1177/08969205241306300 arXiv:https://doi.org/10.1177/08969205241306300

[81] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[82] Duri Long, Jessica Roberts, Brian Magerko, Kenneth Holstein, Daniella DiPaola, and Fred Martin. 2023. AI Literacy: Finding Common Threads between Education, Design, Policy, and Explainability. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 329, 6 pages. doi:10.1145/3544549.3573808

[83] João C. Magalhães and Rik Smit. 2025. Less Hype, More Drama: Open-Ended Technological Inevitability in Journalistic Discourses About AI in the US, The Netherlands, and Brazil. *Digital Journalism* (2025), 1–18. doi:10.1080/21670811.2025.2522281

[84] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. The Social Construction of Generative AI Prompts. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 320, 7 pages. doi:10.1145/3613905.3650947

[85] Andrew Marantz. 2024. Among the A.I. Doomsayers. *The New Yorker* (March 11 2024). https://www.newyorker.com/magazine/2024/03/11/among-the-ai-doomsayers

[86] Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological review* 50, 4 (1943), 370.

[87] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1431, 44 pages.

[88] Jessica McClearn, Rikke Bjerg Jensen, and Reem Talhouk. 2023. Othered, silenced and scapegoated: Understanding the situated security of marginalised populations in lebanon. In *32nd USENIX Security Symposium (USENIX Security 23)*. 4625–4642.

[89] Allison McDonald, Catherine Barwulor, Michelle L Mazurek, Florian Schaub, and Elissa M Redmiles. 2021. "It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online. In *30th USENIX Security Symposium (USENIX Security 21)*. 375–392.

[90] Nora McDonald and Andrea Forte. 2020. The Politics of Privacy Theories: Moving from Norms to Vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376167

[91] Nora McDonald and Andrea Forte. 2021. Powerful Privacy Norms in Social Network Discourse. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 421 (Oct. 2021), 27 pages. doi:10.1145/3479565

[92] Scott McLean, Gemma JM Read, Jason Thompson, Chris Baber, Neville A Stanton, and Paul M Salmon. 2023. The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence* 35, 5 (2023), 649–663.

[93] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing* 1, 4 (2024), 1–45.

[94] Stuart Miller. 1990. Foucault on Discourse and Power. *Theoria: A Journal of Social and Political Theory* 76 (1990), 115–125. http://www.jstor.org/stable/41801502

[95] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* 8, 1 (2021), 141–163.

[96] Niklas Möller, Sven Ove Hansson, and Martin Peterson. 2006. Safety is more than the antonym of risk. *Journal of Applied Philosophy* 23, 4 (2006), 425–432. doi:10.1111/j.1468-5930.2006.00345.x

[97] Linda Monsees, Tobias Liebetrau, Jonathan Luke Austin, Anna Leander, and Swati Srivastava. 2023. Transversal Politics of Big Tech. *International Political*

Sociology 17, 1 (2023). doi:10.1093/ips/olac020

[98] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 599–627. doi:10.1145/3715275.3732039

[99] Lisa P. Nathan, Batya Friedman, Predrag Klasnja, Shaun K. Kane, and Jessica K. Miller. 2008. Envisioning systemic effects on persons and society throughout interactive system design. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems* (Cape Town, South Africa) *(DIS '08)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/1394445.1394446

[100] David B Nieborg, Thomas Poell, and José van Dijck. 2022. Platforms and platformization. *The SAGE handbook of the digital media economy* (2022), 29–49.

[101] Anke Sophia Obendiek. 2025. The politics of tech responsibility: Understanding companies' responsibility as representative claims. *New Media & Society* 27, 6 (2025), 3680–3698. doi:10.1177/14614448241229406 arXiv:https://doi.org/10.1177/14614448241229406

[102] OpenAI. 2025. *A Strong and Safe Start with AI: OpenAI's Teen AI Literacy Blueprint.* White Paper / Policy Blueprint. Available as PDF on OpenAI website.

[103] OpenAI. 2025. Working with 400,000 teachers to shape the future of AI in schools. https://openai.com/global-affairs/aft/. Accessed: 2025-09-08.

[104] OpenAI. 2025. "1972604366160035977" [Tweet]. X (formerly Twitter). https://x.com/OpenAI/status/1972604366160035977?s=20.

[105] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 496–511.

[106] Andrea G. Parker, Laura M. Vardoulakis, Jatin Alla, and Christina N. Harrington. 2025. Participatory AI Considerations for Advancing Racial Health Equity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 803, 24 pages. doi:10.1145/3706598.3713165

[107] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[108] Thomas Poell, David Nieborg, and José van Dijck. 2019. Platformisation. *Internet Policy Review* 8, 4 (2019), 1425. doi:10.14763/2019.4.1425

[109] Penny Powers. 2007. The Philosophical Foundations of Foucaultian Discourse Analysis. *CADAAD Journal* 1, 2 (2007), 18–34. https://ugp.rug.nl/cadaad/article/view/42087

[110] Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1015, 23 pages. doi:10.1145/3706598.3713714

[111] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. doi:10.1145/3531146.3533158

[112] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, et al. 2024. Gaps in the safety evaluation of generative ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1200–1217.

[113] Rainer Rehak. 2025. AI Narrative Breakdown. A Critical Assessment of Power and Promise. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1250–1260. doi:10.1145/3715275.3732083

[114] Janelle Reinelt. 2002. The Politics of Discourse: Performativity Meets Theatricality. *SubStance* 31, 2/3 (2002), 201–215. doi:10.2307/3685486

[115] Anaïs Resseguier and Fabienne Ufert. 2024. AI research ethics is in its infancy: the EU's AI Act can make it a grown-up. *Research Ethics* 20, 2 (2024), 143–155. doi:10.1177/17470161231220946 arXiv:https://doi.org/10.1177/17470161231220946

[116] Veronica A Rivera, Daricia Wilkinson, Aurelia Augusta, Sophie Li, Elissa M Redmiles, and Angelika Strohmayer. 2024. Safer Algorithmically-Mediated Offline Introductions: Harms and Protective Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–43.

[117] Jonathan Roberge, Marius Senneville, and Kevin Morin. 2020. How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society* 7, 1 (2020), 2053951720919968. doi:10.1177/2053951720919968 arXiv:https://doi.org/10.1177/2053951720919968

[118] Camilo Sanchez, Sui Wang, Kaisa Savolainen, Felix Anand Epp, and Antti Salovaara. 2025. Let's Talk Futures: A Literature Review of HCI's Future Orientation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 487, 36 pages. doi:10.1145/3706598.3713759

[119] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction* 2, CSCW (2018), 1–27.

[120] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[121] Emily Setty. 2024. A "post-digital" continuum of young people's experiences of online harms. In *Children, Young People and Online Harms: Conceptualisations, Experiences and Responses*. Springer, 85–111.

[122] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[123] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-centric survey and taxonomy of human-generative-AI interactions. *arXiv preprint arXiv:2310.07127* (2023).

[124] Divyanshu Kumar Singh, Dipto Das, and Bryan Semaan. 2025. The Power of Language: Resisting Western Heteropatriarchal Normative Writing Standards. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 491, 17 pages. doi:10.1145/3706598.3714073

[125] Ananta Soneji, Vaughn Hamilton, Adam Doupé, Allison McDonald, and Elissa M Redmiles. 2024. "I feel physically safe but not politically safe": Understanding the Digital Threats and Safety Practices of OnlyFans Creators. In *33rd USENIX Security Symposium (USENIX Security 24)*. 1–18.

[126] Zahra Stardust, Rosalie Gillett, and Kath Albury. 2023. Surveillance does not equal safety: Police, data and consent on dating apps. *Crime, Media, Culture* 19, 2 (2023), 274–295. doi:10.1177/17416590221111827 Author version. Available at: https://eprints.qut.edu.au/233498/.

[127] Angelika Strohmayer, Rosanna Bellini, and Julia Slupska. 2022. Safety as a grand challenge in pervasive computing: Using feminist epistemologies to shift the paradigm from security to safety. *IEEE Pervasive Computing* 21, 3 (2022), 61–69.

[128] Dag Svanæs. 2025. The Design and Evaluation of a Hands-on AI Literacy Workshop for Politicians. In *STAIG@CHI'25 Accepted Papers*. CHI Conference on Human Factors in Computing Systems. Accepted paper on designing and evaluating AI literacy workshops for politicians.

[129] Tiera Tanksley, Angela D. R. Smith, Saloni Sharma, and Jr Huff, Earl W. 2025. "Ethics is not neutral": Understanding Ethical and Responsible AI Design from the Lenses of Black Youth. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 200, 20 pages. doi:10.1145/3706598.3713510

[130] Anna-Lena Theus. 2023. Striving for affirmative algorithmic futures: How the social sciences can promote more equitable and just algorithmic system design. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 558–568.

[131] Olena Trofymenko, Yuliia Prokop, Natalia Loginova, and Alexander Zadereyko. 2021. Taxonomy of Chatbots. In *Proceedings of the II International Scientific and Practical Conference "Intellectual Systems and Information Technologies" (ISIT 2021)*. CEUR Workshop Proceedings, Odesa, Ukraine, September 13–19. https://ceur-ws.org/Vol-XXX/ Available under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[132] Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. "Ownership, Not Just Happy Talk": Co-Designing a Participatory Large Language Model for Journalism. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)* (Athens, Greece). ACM, 12. doi:10.1145/3715275.3732198

[133] United Nations. 2023. Taxonomy of Human Rights Risks Connected to Generative AI. United Nations Human Rights Office of the High Commissioner. https://tinyurl.com/3mffabyf.

[134] Risto Uuk, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl. 2024. A taxonomy of systemic risks from general-purpose AI. *arXiv preprint arXiv:2412.07780* (2024).

[135] Teun A. van Dijk. 2017. *Discourse and Power* (1st ed.). Bloomsbury Publishing, New York. https://www.bloomsbury.com/us/discourse-and-power-9781137072993/ Ebook (PDF), Red Globe Press imprint.

[136] Johanna Velander, Nuno Otero, and Marcelo Milrad. 2024. What is critical (about) AI literacy? Exploring conceptualizations present in AI literacy discourse. In *Framing Futures in Postdigital Education: Critical Concepts for Data-driven Practices*, Anders Buch, Ylva Lindberg, and Tessy Cerrato-Pargmant (Eds.). Springer, Cham, Switzerland. https://gala.gre.ac.uk/id/eprint/47286 In Press.

[137] Peter Verdegem. 2022. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. *AI & Society* (April 2022), 1–11. doi:10.1007/s00146-022-01437-8 Epub ahead of print.

[138] Sandra Wachter. 2023. Limitations and loopholes in the EU AI Act and AI Liability Directives: what this means for the European Union, the United States, and beyond. *Yale JL & Tech.* 26 (2023), 671.

[139] Weili Wang and John Downey. 0. Mapping the sociotechnical imaginaries of generative AI in UK, US, Chinese and Indian newspapers. *Public Understanding of Science* 0, 0 (0), 09636625251328518. doi:10.1177/09636625251328518 arXiv:https://doi.org/10.1177/09636625251328518 PMID: 40219721.

[140] Weili Wang, John Downey, and Fan Yang. 0. AI anxiety? Comparing the sociotechnical imaginaries of artificial intelligence in UK, Chinese and Indian newspapers. *Global Media and China* 0, 0 (0), 20594364231196547. doi:10.1177/20594364231196547 arXiv:https://doi.org/10.1177/20594364231196547

[141] Max Weber. 1978. *Economy and Society.* University of California Press, Berkeley. Original work published 1922.

[142] Max Weber. 2015. Politics as Vocation. In *Weber's Rationalism and Modern Society*, Tony Waters and Dagmar Waters (Eds.). Palgrave Macmillan, New York, 129–198. Original work published 1919.

[143] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

[144] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency.* 214–229.

[145] Martin Winkel. 2025. Controlling the Uncontrollable: The Public Discourse on Artificial Intelligence Between the Positions of Social and Technological Determinism. *AI & Society* 40 (2025), 1947–1959. doi:10.1007/s00146-024-01979-z

[146] Richmond Y. Wong and Steven J. Jackson. 2015. Wireless Visions: Infrastructure, Imagination, and US Spectrum Policy. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15).* Association for Computing Machinery, New York, NY, USA, 105–115. doi:10.1145/2675133.2675229

[147] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 145 (April 2023), 27 pages. doi:10.1145/3579621

[148] Takuya Yokota and Yuri Nakao. 2025. Stakeholder Participation in AI Auditing: Challenges and Future Directions. In *Fujitsu Research Publications.* Fujitsu Limited, Kawasaki-city, Japan. Position paper discussing multi-stakeholder involvement in AI development and auditing.

[149] Aurora Zhang and Anette Hosoi. 2024. Structural Interventions and the Dynamics of Inequality. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 1014–1030.

[150] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 13, 17 pages. doi:10.1145/3706598.3713429

A Appendix: Texts Cited from Corpus

ID	Link
15	https://openai.com/index/our-approach-to-ai-safety/
29	https://www.anthropic.com/news/our-approach-to-understanding-and-addressing-ai-harms
4	https://openai.com/business-data/
19	https://openai.com/index/openai-safety-update/
43	https://www.anthropic.com/news/reflections-on-our-responsible-scaling-policy
45	https://www.anthropic.com/research/sabotage-evaluations
3	https://openai.com/index/openai-board-forms-safety-and-security-committee/
63	https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf
68	https://policies.google.com/terms/generative-ai/use-policy
47	https://www.anthropic.com/news/the-case-for-targeted-regulation
35	https://www.anthropic.com/news/core-views-on-ai-safety
58	https://blog.google/technology/ai/why-we-focus-on-ai-and-to-what-end/
44	https://www.anthropic.com/news/anthropics-responsible-scaling-policy
116	https://deepmind.google/discover/blog/taking-a-responsible-path-to-agi/
34	https://www.anthropic.com/news/constitutional-classifiers
33	https://www.anthropic.com/research/clio
121	https://openai.com/charter/
40	https://www.anthropic.com/news/national-security-expert-richard-fontaine-appointed-to-anthropic-s-long-term-benefit-trust
27	https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations
36	https://www.anthropic.com
9	https://openai.com/safety/how-we-think-about-safety-alignment/
21	https://openai.com/transparency-and-content-moderation/
126	https://blog.google/outreach-initiatives/public-policy/google-us-ai-action-plan-comments/
121	https://openai.com/charter/