How Generative AI Empowers Attackers and Defenders Across the Trust & Safety Landscape

Patrick Gage Kelley
patrickgage@acm.org
Google
USA

Steven Rousso-Schindler
stevenrs@gmail.com
CSU Long Beach
USA

Renee Shelby
reneeshelby@google.com
Google Research
USA

Kurt Thomas
kurtthomas@google.com
Google
USA

Allison Woodruff
woodruff@acm.org
Google
USA

Abstract

Generative AI (GenAI) is a powerful technology poised to reshape Trust & Safety. While misuse by attackers is a growing concern, its defensive capacity remains underexplored. This paper examines these effects through a qualitative study with 43 Trust & Safety experts across five domains: child safety, election integrity, hate and harassment, scams, and violent extremism. Our findings characterize a landscape in which GenAI empowers both attackers and defenders. GenAI dramatically increases the scale and speed of attacks, lowering the barrier to entry for creating harmful content, including sophisticated propaganda and deepfakes. Conversely, defenders envision leveraging GenAI to detect and mitigate harmful content at scale, conduct investigations, deploy persuasive counternarratives, improve moderator wellbeing, and offer user support. This work provides a strategic framework for understanding GenAI's impact on Trust & Safety and charts a path for its responsible use in creating safer online environments.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics → Computing / technology policy**; • **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Artificial intelligence**.

Keywords

generative AI, frontier AI, Trust & Safety, child safety, election integrity, hate and harassment, online scams, violent extremism

ACM Reference Format:
Patrick Gage Kelley, Steven Rousso-Schindler, Renee Shelby, Kurt Thomas, and Allison Woodruff. 2026. How Generative AI Empowers Attackers and Defenders Across the Trust & Safety Landscape. 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, 21 pages. https://doi.org/10.1145/3772318.3791363

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791363

Content Warning: This paper discusses topics that some readers may find distressing, including child sexual abuse material (CSAM), hate speech, harassment, and violent extremism.

1 Introduction

Recent advances in artificial intelligence have substantially raised expectations that it will transform work and society in areas such as economic productivity, scientific progress, education, and quality of life, among others [85, 87, 137]. Many of these expectations stem from the newest GenAI systems, and their general-purpose nature, which allows them to be used for a wide variety of tasks across nearly every domain.

As with many areas, this technological shift is poised to broadly reshape the landscape of Trust & Safety, a field that has gained increased visibility in the CHI community in recent years [22, 82]. Trust & Safety encompasses the organizations and disciplines focused on protecting the internet from myriad harmful activities and content, ranging from child safety to misinformation, through content moderation, platform policies, signal gathering, and investigations. The Trust & Safety field is characterized by an ongoing interplay between *attackers* (e.g., scammers, harassers) and *defenders* (e.g., content moderation teams, child safety hotlines, law enforcement). GenAI may fundamentally change the threat landscape across Trust & Safety—who attackers are, who they target, how they operate, and the nature of their attacks. Concurrently, GenAI offers significant, underdeveloped opportunities for defenders to restructure their operations to counter both novel GenAI-enabled attacks and longstanding ones [28]. While prior research has focused on technical safeguards to minimize misuse of AI models and systems [3, 95, 113, 123, 133, 135], these studies fall short of detailing how the uneasy sociotechnical balance between attackers and defenders will evolve alongside GenAI.

We advance HCI's contribution to Trust & Safety by exploring how GenAI impacts the field, centering the perspectives of Trust & Safety experts on the front lines of fighting harmful content online. Drawing on participatory methods from HCI, we conducted six-hour participatory research workshops for five different Trust & Safety domains with a total of 43 expert participants from Asia, Europe, and North America. Our contributions include:

- We present a **novel qualitative study of experts in five Trust & Safety domains**—child safety, election integrity, hate and harassment, scams, and violent extremism—exploring how GenAI currently affects their domains, and how they expect it to affect them in the future.
- We describe a **dominant narrative that GenAI empowers both attackers and defenders**, a view shared in common by participants across all Trust & Safety domains studied.
- We describe the ways in which participants expect **escalating competition between attackers and defenders** to unfold, with GenAI initially enabling attacks that outpace existing defense systems but subsequently enabling advanced defensive capabilities. At a high level, this view is shared in common across domains, but the specifics vary in different domains.
- We identify **structural differences in different domains**, such as the nature of harm and the scale of attacker operations. Based on these differences, we draw practical implications for Trust & Safety operations by highlighting the need for domain-specific interventions.
- Building on our participants' insights, we describe **the imperative, challenges, and opportunities for Trust & Safety organizations to transform their operations** to meet the novel threats posed by GenAI, and also, to fully realize the advantages of GenAI.

In the remainder of the paper, we review relevant background, describe our methodology, present our findings, and discuss tangible paths for creating safer online environments.

2 Background

To situate our study of GenAI's impact on Trust & Safety, we review the technical foundations of generative models, Responsible AI and Trust & Safety, Trust & Safety domains, and emerging literature on GenAI's effects in this area.

2.1 GenAI Models

The most recent leap forward in artificial intelligence systems has been a set of technologies broadly referred to as GenAI, which includes large language models (LLMs) and other frontier models. These GenAI systems generate realistic or high-quality synthetic text, images, code, audio, and video, typically in response to a set of instructions called a prompt. Technologically, these systems are driven by an advancement in system architecture, the attention-based transformer model [132], which allows for training extremely large models on vast datasets using highly parallelized modern chip architectures. GenAI systems are marked by three key characteristics: (1) applicability to generalized rather than specialized use cases; (2) production of original content that is often indistinguishable from human-created content; and (3) intuitive and accessible interfaces [20]. These systems are deployed as both proprietary models, typically equipped with safety guardrails by their developers, and as open-source models, which offer greater customizability and can be run with fewer inherent restrictions.

2.2 Responsible AI and Trust & Safety

Responsible AI (RAI) is an umbrella term for research and practices aimed at developing safe, reliable, and ethical AI systems in a manner that aligns with societal values. Covering the entire system lifecycle, RAI is often structured around key principles [55], such as *fairness*, to mitigate algorithmic bias; *accountability*, to establish clear lines of responsibility; and *transparency*, to ensure model decisions are understandable [34]. This upstream approach to technology development is enacted through governance structures and best practices for AI design, development, and deployment [96, 115].

This applied work translates high-level RAI principles into concrete technical safeguards, focusing on preventing sociotechnical harms [113, 135] by ensuring the robustness, reliability, and security of AI models and systems [3, 95]. It addresses vulnerabilities to prevent both direct misuse and harm from benign use [123]. This work involves a range of activities, including identifying social and ethical risks [101]; implementing pre-training interventions, such as improving the data the model learns from [123]; applying post-training mitigations to prevent undesirable behaviors like sycophancy or goal misinterpretation [94]; and ensuring the fairness and robustness of tools used for content moderation, such as multimodal classifiers that detect hate speech [107] or other policy-violating content [41]. The field also focuses on mitigating risks such as "jailbreaking," where users circumvent safety controls with adversarial prompts, and improving scaled evaluation techniques [133].

Complementing these safeguards are the operational practices of Trust & Safety, which strives "to promote a safer and more trustworthy internet" [29]. Historically, the field operated under a reactive model of incident response, such as removing user-flagged content or responding to legal requests after harm occurred [24, 74]; however, Trust & Safety is shifting toward a proactive strategy to embed safety into the design and architecture of a system [131]. Trust & Safety roles vary significantly depending on a company's technology, user base, or the particular harms on which they focus. The Trust & Safety Professional Association (TPSA) highlights diverse roles spanning policy design, threat analysis, and content moderation—many of which are similar to Responsible AI jobs [100]. Professionals in these roles engage in a broad spectrum of activities, such as developing platform rules, supporting users, and moderating content [13]. The central aim is to mitigate risks and safeguard users from malicious actors, thereby enhancing user experience, reinforcing public trust, and meeting the complex legal requirements of various international frameworks.

2.3 Trust & Safety Domains

Defense against online harms is not the sole responsibility of technology platforms but is instead managed by a diverse, multi-stakeholder ecosystem. This network includes platform Trust & Safety teams, civil society organizations, government agencies, academic researchers studying emerging threats, and industry bodies that facilitate cross-company collaboration [29]. The resilience of this ecosystem relies on effective coordination, as no single actor can address these complex challenges alone.

Our work focuses on five key domains where this ecosystem confronts persistent, high-severity threats: child safety, election

integrity, hate and harassment, scams, and violent extremism. We briefly describe these domains and some of the existing technologies used for detection and removal. A more exhaustive discussion of content moderation can be found in Singhal et al. [117] along with some of the professional challenges in Moran et al. [82].

Child safety is a foundational domain within Trust & Safety, encompassing a wide spectrum of risks to minors, including grooming, cyberbullying, and exposure to age-inappropriate content. Among these harms, child sexual abuse material (CSAM) represents one of the most severe and legally unambiguous threats. CSAM involves the creation and distribution of real, photorealistic, or non-photorealistic sexual imagery of individuals under the age of 18. According to the International Center for Missing and Exploited Children (ICMEC), CSAM is illegal globally in all but 10 countries as of 2023 [52], though definitions differ by jurisdiction. Historically, a relatively limited number of images—estimated at several hundred thousand—has circulated repeatedly [80]. Accordingly, technology platforms rely on perceptual hashes to match and report previously identified CSAM, with popular tools including PhotoDNA [63] and Google's Content Safety API [40]. In the United States, the National Center for Missing and Exploited Children (NCMEC) serves as a clearinghouse for sharing CSAM hashes between platforms. NCMEC also receives reports of any detected content, with over 20 million such reports in 2024 [86]. Longstanding challenges in this domain include the growing volume of content, the emotional toll on reviewers, and the risk of new material evading hashed-based detection [21]. In 2023, NCMEC had its first "million report day," due to a single viral meme which was manageable only with automated clustering, raising concerns that GenAI could enable a similar volume of unique images and overwhelm existing response systems [44, 125].

Election integrity covers threats such as false claims about voting procedures and fraud, intimidating or dissuading voters, and fabricated content related to candidates (e.g., deepfakes). Concerns have surfaced regarding the use of GenAI by attackers, although some analyses suggest the harms to date have been more subtle than anticipated [83, 84, 108]. Existing protections include fact checking partnerships, media literacy campaigns, and disclosure requirements for state-sponsored media.

Hate and harassment refers to a broad spectrum of toxic content (e.g., hate speech, trolling, sexual harassment) and maliciously leaked content (e.g., doxxing, non-consensual explicit imagery) [126]. Formal definitions of hate speech reflect just part of this spectrum, resulting in a fractured ecosystem of policies which is challenging for both moderators and users [99, 112, 121]. Approximately 48% of people globally report experiencing some form of online hate and harassment [126]. The scale, decentralization, and polymorphism of this content require solutions beyond hashing, most often platform-specific content classifiers (e.g., the Perspective API [30, 138]) and user reporting.

Scams are one of the broadest online threats, costing victims billions in losses annually [4, 127]. Examples include counterfeit products [65], cryptocurrency giveaways [66, 69], and "pig butchering" attacks [1, 90], among others. The profit incentive for scammers has led to an arms race in detection and evasion, with increasingly sophisticated scam infrastructure and content [140]. Common detection strategies include content-based classifiers, account-based classifiers [139], and browser-based warnings [31], most of which are custom to each platform.

Violent extremism (VE) glorifies, encourages, or facilitates violence to support ideological goals [7, 79]. Global definitions of VE and the specific entities it encompasses vary by country. Similar to CSAM, platforms rely on hash-based detection, with the Global Internet Forum to Counter Terrorism (GIFCT) acting as a hash-sharing clearinghouse. As of early 2024, GIFCT's database contained over 2.3 million hashes covering 408,000 distinct images, videos, and texts [130]. Efforts to leverage GenAI for violent extremist and terrorist purposes have already been documented by several organizations [72, 116, 124].

2.4 The Emerging Impact of GenAI on Trust & Safety

A growing body of research and news articles has posited the risks GenAI poses in creating and distributing harmful content. Risks cited include promoting conspiracy theories [50], producing highly-believable fabricated media [84], increasing the quantity and quality of phishing scams [49], generating synthetic, non-consensual sexually explicit imagery or CSAM [37, 56, 134], or otherwise coercing users [9, 73]. These risks are exacerbated by the possibility of attackers circumventing AI safety controls to produce harmful content [81, 133] and the inherent challenge of defining safety controls robustly [76, 99]. Our work expands on these perspectives by engaging experts across the Trust & Safety ecosystem on the specific nuances of how GenAI will influence each of the five domains in our study.

Comparatively less attention has been paid by researchers and the media to the potential benefits of GenAI for improving defenses against harmful content. Prior studies have explored using AI to help debunk conspiracy theories [104], better detect scam content [10], or scale the detection of hate and harassment, violent extremism, and election integrity threats [28, 121, 128]. Our work builds on these findings by exploring these potential applications from the perspectives of experts, identifying the opportunities and gaps they believe GenAI could help fill. This includes countering pre-existing threats, as well as new threats stemming from GenAI. While a significant body of literature exists for improving content moderation within specific Trust & Safety domains [117], far less research has examined the cross-cutting challenges and opportunities that connect them. Furthermore, much of the existing technical work focuses on automated detection, with a dearth of research directly engaging the human experts who operate within these complex sociotechnical systems [82].

3 Methodology

To examine the effects of GenAI on Trust & Safety, we conducted participatory research workshops in Asia, Europe, and North America with experts from five domains: child safety, election integrity, hate and harassment, scams, and violent extremism. We draw on and extend the methodology presented in Woodruff et al. [137], which studied the impact of GenAI on knowledge workers. Our research examines the following questions:

- How is GenAI already impacting Trust & Safety domains?
- How is GenAI likely to impact Trust & Safety domains in the future?
- How does GenAI's impact on Trust & Safety domains compare or interact with other changes in the field?

3.1 Participant Recruitment

As Trust & Safety covers a wide range of digital safety harms, we selected five domains that have been highlighted in the Trust & Safety literature as top priorities [26]. In consultation with Trust & Safety experts, we confirmed these domains—child safety, election integrity, hate and harassment, scams, and violent extremism—vary in their harm profile, scope and scale, and regulatory environment, allowing us to elicit diverse perspectives across Trust & Safety issues. Furthermore, these domains are likely to be substantially affected by GenAI [28]. Considering multiple domains allowed us to interrogate existing operational and regulatory assumptions, which often take a one-size-fits-all approach to Trust & Safety harms management and mitigation. For more information about the domains, see Table 1.

Following prior HCI and participatory design work examining how experts sensemake about novel technologies [57, 78, 111] and the impacts of technologies in their domains [45, 58, 89], we engaged Trust & Safety workers as expert participants. Engaging experts is valuable because of their domain-specific knowledge, which offers essential insights regarding on-the-ground needs and norms [70, 71, 93, 101] and situated perspectives towards potential sociotechnical harms [114].

Accordingly, we recruited 8-10 workshop participants from each of the five domains, for 43 in total. For each workshop, we sought a mix of participants from civil society, industry, and academia. Within their domains, participants represented perspectives in areas such as education, human rights, journalism, law, law enforcement, machine learning, mediation, policy, and social science. Each participant received a personal invitation as a known expert in their field. We recruited participants in two ways: through mutual or personal connections or expressions of interest, and by identifying and contacting them based on relevant publicly visible expertise.

In this expertise-centered approach, participants and researchers collaboratively share and build upon each other's knowledge to reach a deeper understanding of a topic and envision solutions sensitive to the needs of multiple stakeholders (see Section 3.2). In such an approach, it is neither expected nor necessary that experts fully grasp the underlying technical implementation of a particular technology to contribute meaningfully to the discussion of its impacts in their domains [46], but rather, that they have expertise relevant to the research objectives. Thus, we prioritized recruiting participants with deep domain knowledge and on-the-ground experience rather than requiring specific technical knowledge of GenAI (although in fact many of the participants did have extensive experience with GenAI, especially regarding how it is used in their domains). Part of our role as researchers was to provide participants sufficient information about GenAI for them to relevantly engage, and it was also our role to gently address any potential misconceptions during discussion. Accordingly, we led a teaching intervention early in each workshop to bolster their understanding of GenAI (see Section 3.3) and continued collaborative discussion of technological capabilities and characteristics throughout the sessions.

While our institution does not have an Institutional Review Board (IRB), we adhere to similarly strict standards. We received informed consent from all participants using a standard form, which included consent for a videographer to collect high-quality video and audio recordings. All participants were compensated at industry standard, receiving an equivalent amount in their local currency. A few participants who could not receive compensation directly due to institutional policies were offered the opportunity to direct the funds as a charitable donation. Additionally, travel reimbursements were available to most participants. For a summary of participants in each location, see Appendix A, Table 4.[1]

3.2 Workshops as Research Method

We conducted participatory workshops to engage with communities of practice representing five areas of Trust & Safety work. As a method, participatory workshops offer a structured way for researchers to gather "reliable and valid data about [a] domain ... regarding forward-oriented processes, such as organisational change" [92, p. 73]. This approach enables researchers to engage stakeholders to collaboratively explore their situated knowledge [5, 48], thereby grounding the inquiry in specific real-world contexts [105]. In HCI, participatory workshops are a well-established methodology [64, 105], particularly well-suited for engaging participants on emerging topics and capturing the "real-timeness" of a phenomenon, meaning actively engaging with the problem as it unfolds [27]. They have been used to convene participants to reflect on domain-specific challenges [118, 129], discuss technologies relevant to their work [2], identify domain-specific harms [114], and engage in speculative design [98, 106]. We adopted this methodology specifically because of its alignment with critical theoretical traditions in HCI (e.g., feminist, queer, and critical race perspectives) that recognize how technologies are differently experienced in ways shaped by situated norms and extant power dynamics [8, 61, 103, 109]. This approach allowed us to: (1) center the direct, lived experiences of people affected by technology [8, 103]; (2) examine technology as a driver of social change [68]—in our case, understanding the stages of GenAI adoption in Trust & Safety work and exploring participatory views on potential sociotechnical and organizational interventions; and (3) employ research tactics designed to dismantle traditional hierarchies between researchers and participants [48, 97, 120]. To achieve these critical and contextual objectives, we designed our workshops to foster collaborative reflection, envisioning, and discussion, incorporating probes [11, 16, 19, 35, 42] and provocations to inform future agendas and policies around GenAI in Trust & Safety.

3.3 Workshop Structure

We held a total of five workshops, one per domain, between March 2024 and January 2025. We selected locations that are centers of activity for the respective domains, as well as being in different

[1] In keeping with participant privacy practice discussed in Bellini et al., we provide limited detail about participants in order to minimize risk of identification [12].

Domain	Identifier	Brief Overview
Child Safety	*C*	*Ensure child safety online. The discussion largely focused on preventing the creation and sharing of child sexual abuse material (CSAM) online.*
Election Integrity	*E*	*Promote the integrity of elections, considering both misinformation and disinformation. The discussion encompassed civics integrity as well.*
Hate & Harassment	*H*	*Counter online hate & harassment with approaches such as content moderation and content policy.*
Scams	*S*	*Conduct anti-scam operations, to combat a wide range of online scams, such as romance scams or videoconference scams.*
Violent Extremism	*V*	*Counter violent extremism and terrorist activities which are facilitated by online content and behavior.*

Table 1: Overview of each domain. In the Findings section, Identifiers are used to attribute quotations to domains.

Activity	*Time*
Arrival, Breakfast	
Welcome, Introductions	*30 minutes*
Memorable Experience Cards Activity	*50 minutes*
Mid-Morning Break	*10 minutes*
GenAI Teaching and Q/A	*45 minutes*
Lunch Break	*45 minutes*
Change Cards Activity	*60 minutes*
Attackers Discussion	*30 minutes*
Mid-Afternoon Break	*15 minutes*
Futuring Stories Activity	*60 minutes*
Wrap-Up	*15 minutes*

Table 2: Agenda for participatory research workshops. Times are approximate.

regions to represent a range of international perspectives. Furthermore, we held most workshops proximate to other large conferences relevant to the respective domains, in order to reduce travel burden for participants. Accordingly, we held workshops in Asia, Europe, and North America, with participants from a total of 15 countries (for specific cities and domains, see Appendix A, Table 4). Each six-hour workshop was held in person at a local Google office, and participants were aware of Google's involvement in the study. Each workshop was moderated by two or three researchers. An additional researcher, a visual ethnographer, attended each workshop in the role of videographer. The workshop agenda is shown in Table 2.

In designing this six-hour agenda, we extended the three-hour workshop format detailed in Woodruff et al.'s study of knowledge workers [137]. Our workshops retained two activities previously used successfully with knowledge workers: the GenAI teaching and Q/A, and the change cards activity [137]. We customized these activities to the Trust & Safety context, leveraging our domain expertise, and we added three new activities using similar probes: the memorable experience cards activity, the attackers discussion, and the futuring stories activity. We piloted these new activities in the first workshop, and iteratively revised the futuring stories activity for the subsequent workshops as described below. The memorable experience cards activity and the attackers discussion required no modification.

We began by welcoming participants and setting expectations for the day. After a round of introductions, we facilitated a discussion about participants' memorable experiences with GenAI. For this purpose, we introduced a probe in the form of a small, physical card: a *memorable experience card* (Appendix B, Figure 1). These cards encourage participants to reflect on any memorable experiences they have had with GenAI, whether related to their domain or not. Each participant spent about 10 minutes filling out their individual card(s). Participants then shared one or more of their memorable experience cards in a group discussion which we facilitated, allowing time for participants to react to each other's experiences. This activity introduced participants' orientation to and impressions of GenAI and laid the groundwork for future discussion. We then held a short break.

Next, a researcher from our team gave a presentation about GenAI, followed by a question and answer session, to provide participants with a shared understanding of this technology. The presentation focused on the key characteristics of GenAI as well as

providing additional historical context (Appendix C). This presentation was followed by a meal break.

We then facilitated a discussion of potential upcoming changes to Trust & Safety. We began with a probe introduced in Woodruff et al. [137]: a large, physical card called a *change card*, which invites participants to share significant changes (related to GenAI, or not) that could take place in their domain in the next one to two years (Appendix B, Figure 1). As described in Woodruff et al. [137], change cards support reflection and "capture thoughts that participants [can] revisit and build on in collaborative discourse." Each participant spent about 10 minutes filling out their individual change card(s). Participants then shared and discussed the content of their change cards in an open-ended group discussion which we facilitated. Following this activity, we transitioned to an open-ended discussion of how participants believe attackers may use GenAI in their domains. We then held another break for refreshments.

Finally, inspired by futuring approaches [59, 60, 88], we invited participants to write an optimistic story set two years in the future about GenAI being used to effectively mitigate harm in their domains.[2] We encouraged participants to include classic storytelling elements such as setting, characters, plot, and/or conflict. We gave verbal instructions and provided each participant a large, physical *futuring card* with the prompt, on which they wrote their stories (Appendix B, Figure 1). Participants spent about 15 minutes working individually, and we included a checkpoint about halfway through this time to ask if participants had any questions or would like any support with their stories. After the individual work concluded, participants took turns reading their stories aloud in a facilitated discussion that lasted approximately 45 minutes, and we then concluded the session. Throughout the workshops we attended carefully to participant dynamics and supported collaborative discussion amongst all attendees.

3.4 Data Preparation and Analysis

The workshops produced three types of data: verbatim transcripts of the workshop discussions, participant-created artifacts (cards and stories), and researcher-produced synthesized video "clip reels" prepared by the visual ethnographer. Initial verbatim transcripts were created using an automated transcription service, and subsequently corrected for accuracy by the research team. This correction process involved reviewing the original recording while simultaneously reading and making necessary adjustments to the transcript. Identifying participant details were replaced with assigned pseudonyms (e.g., , C1, C2) to maintain confidentiality. Participant-created artifacts, including 125 memorable experience cards, 99 change cards, and 28 futuring stories, were digitally scanned for analysis. For each workshop, the visual ethnographer created a video "clip reel" distilling the six-hour session into approximately 30 minutes of content highlighting participants' strong opinions and compelling stories aligned with the research questions.

We undertook a reflexive thematic analysis [17, 18] of this corpus, to inductively build themes centered on the impact of GenAI in Trust & Safety domains. The analysis involved deep and prolonged data immersion as well as discussion during meetings. Four members of the research team engaged in initial, independent open coding of the textual and video data, with concurrent analytical memoing to capture initial emergent themes [15]. A designated researcher focused on open-coding the artifacts and integrating their findings into the shared analytical discussions. Throughout our analysis, the four researchers, including the visual ethnographer, held regular meetings to iteratively cross-compare codes and interpretations to reach consensus.[3] The process continued until theoretical saturation of the data was reached, allowing us to finalize our main themes as well as create domain-specific thematic composites [25, 136]. The research team is composed of researchers with direct experience in both GenAI and the Trust & Safety domains studied, as well as related areas such as responsible AI development and AI governance. This experience allowed us to better interpret participants' comments, and assess alignment with consensus technical views of GenAI to validate the findings. Further, our experience allowed us to translate participants' insights about on-the-ground challenges and needs into actionable findings for Trust & Safety operations. Although the research team drew on our professional experiences, we engaged in a shared reflexive practice, continually discussing and interrogating the assumptions that arose during the analysis.

3.5 Limitations

We invite readers to note several limitations of this study. Given that the research was conducted in only four international locations with single expert cohorts per domain, these findings offer a rich, contextually bounded perspective on situated expert knowledge but should not be taken as more broadly representative. In addition, the design of the workshops may have introduced a potential framing effect on participant commentary. This was counterbalanced, however, by prioritizing open, participant-led discussion and the participants' own pre-existing expert opinions. Finally, it is important to acknowledge that both the landscape of GenAI and expert understanding of it are changing rapidly. Therefore the expert consensus captured here reflects a snapshot in time, and views in Trust & Safety domains may shift as the technology matures.

4 Findings

Participants from all domains shared that GenAI can empower both attackers and defenders. Attackers can use GenAI for harmful purposes online. At the same time, defenders can use GenAI to implement safeguards against harm, whether it is perpetrated by GenAI or other means. For example, GenAI can enable both scam and anti-scam efforts.

> "I think scams will continue to proliferate [within the next two to five years]... They get better, we get better. And the arms race keeps progressing." –*S6*[4]

[2] In the first workshop, we piloted a variant of this activity, seeded with post-it notes. Based on learnings from this activity, we settled on the final version described here for the four remaining workshops.

[3] Consistent with practices commonly used for rich analysis of group interviews as described in McDonald et al. [75], we focused on achieving researcher consensus through nuanced discussion rather than measuring inter-rater reliability.

[4] Throughout the paper, we assign participants pseudonyms that begin with the letter associated with their domain as shown in Table 1, i.e., *C* (Child Safety), *E* (Election Integrity), *H* (Hate & Harassment), *S* (Scams), and *V* (Violent Extremism). In some cases, we have lightly cleaned the quotes for readability, e.g., to remove inconsequential false starts or to remove filler words such as "like" or "um."

> "Whilst clearly there's a huge threat in terms of the terrorist use of these technologies, there's also a significant opportunity in terms of using generative AI to understand, intervene, communicate, and so on and so forth [to counter terrorism]. So, incredibly excited by the benefits that generative AI will bring." *–V2*

> "Bad actors will learn and will be able to use these tools to do whatever they want to do. But that's just what the tool is. It can be used for bad things as well as good things, and the good things are right next to it... If [bad actors] use generative AI to generate propaganda, that's something that can be countered by the opposite side. Like anti-extremists saying, 'Okay, we can use the same tool [as extremists], but make good content.'" *–C3*

In this section, we describe participants' perspectives on the co-evolution of attackers and defenders in their use of GenAI. We begin by describing how GenAI can empower attackers, highlighting views held in common across domains. Next, we describe how GenAI can empower defenders, again highlighting views held in common across domains. We illustrate participants' concerns that GenAI will initially overwhelm existing, overtaxed response systems and that the defense ecosystem is reacting too slowly, as well as their optimism that GenAI can eventually be leveraged to counter both preexisting threats and new attacks facilitated by GenAI. Finally, we turn to unique perspectives within specific domains.

4.1 How Generative AI Empowers Attackers

Participants in all domains were concerned about "all the innovative ways that bad actors are using AI" and how GenAI can "empower cyber criminals." Participants described memorable experiences with GenAI-generated harmful content, highlighting current, active, damaging uses of GenAI.

> "I've seen obviously some sets of [CSAM] images or videos multiple times, as many of us probably have, and so I would say I know the faces of some of the victims of course... [I remember] the first time I think I saw an AI-generated image that was supposed to be photorealistic. But actually this one... they called it a 'new set of images,' of a victim that had been abused years ago, and she's an adult now. Probably most of us know her actually. And I just saw these images and they looked almost photorealistic. This is when I thought, 'Okay, this is an even worse step than just generating CSAM or other horrible content with AI or with other tools.' And actually making new content of victims that had been abused [years ago] is a whole other level that was very different for me in that moment." *–C3*

They anticipated even more harm in the future, given the sophistication and rapid rate of adoption of GenAI tools. They emphasized that existing Trust & Safety operations are not equipped to meet these "alarming" and "terrifying" new threats. In the remainder of this section, we describe how scale and speed, lowered barriers to entry, enhanced messaging, degradation of the information environment, and harmful personal relationships with chatbots or AI companions empower attackers.

4.1.1 Scale and Speed. Participants explained that attackers are using GenAI to scale up harmful online content and behavior in ways that outpace existing defense mechanisms. Participants highlighted that GenAI dramatically increases the volume and speeds up the production of harmful content in every Trust & Safety domain studied, and as a result, "all [harmful] content is just exploding in volume."

> "AI CSAM has dramatically risen as a new category..." *–C9 (Memorable Experience Card)*

> "... not being able to keep up with the scale of content moderation, because now it's just so much easier, faster, cheaper to do any of it. It's so much cheaper than renting a troll farm or something. And you can make it a lot more targeted and personalized... Content moderation was already really, really hard and something that didn't always get enough time or resources. And it just got so much harder." *–H2*

Taking scams as an example, using GenAI can be easier and faster than perpetrating them manually. For example, if a romance scam involves sending text messages over a prolonged period, GenAI can create and send those messages. Consequently, where a single individual might previously have spent time messaging a small number of people, they can now oversee dashboards of GenAI chatbots sending messages to a much larger number. This pattern, where a single actor can multiply their impact, was a common concern across all domains.

Accordingly, participants emphasized the strain GenAI places on existing content review, legal, and investigative systems. They saw GenAI attacks as taxing an already under-resourced and underprioritized Trust & Safety ecosystem, with some describing the status quo as a failed system. In child safety, participants observed that existing mechanisms such as hashlists would need to evolve both technically and operationally to defend against new challenges posed by GenAI. Furthermore, participants expressed concern that the defender ecosystem is adapting too slowly to challenges raised by rapidly evolving GenAI technologies. At the same time, they worried that distant future threats receive disproportionate attention at the expense of urgent, ongoing problems.

> "We're already seeing increasing workload, all of us. Hugely increasing. Not like ten percent a year, more like sometimes a couple of hundred percent more a year, in reports that we get, in content that we have to assess and work on. On the other hand, our resources are not exploding like that. They're more stagnating." *–C3*

> "What we have seen pretty consistently [so far] is that it's the use of broadly shared open-source models that are manifesting in these kinds of environments. They're still mostly diffusion models... When I think about detection, one of the things that I'm concerned about is [attackers will] eventually find some newer, more exciting architecture than diffusion models to produce these [CSAM] materials, and then all of the detection technologies we've built will now be out of date and will need to be rebuilt in order to address that problem." *–C10*

> "Oftentimes our discussions of what AI can do end up focusing too much on the speculative... I think I focus too much on, 'Hey, what's gonna happen in the next few years?' when the harm that is happening right now is very apparent..." *–V4*

> "These are actually really hard problems and what's really concerning is that there's a lot of time and effort and resources and money just being spent in trying to figure out what we're gonna build in the next 10 years, and how we're going to get AGI or what that will look like, versus, 'Let's solve the problems that we have today.' There's a lot of them, and there just isn't enough time or focus or just money being spent on that." *–H2*

4.1.2 Lowered Barriers to Entry. Participants expressed significant concern that the widespread availability and ease of use of GenAI tools have lowered the barrier to entry for creating and propagating harm. Readily available GenAI tools require relatively little technical skill, making it easier to create harmful content and widening the range of potential bad actors. For example, participants raised concerns about the rise of AI nudification apps and their use by teens. Similarly, just as specialized software toolkits previously enabled more people to engage in scams, participants anticipate that GenAI tools will pave the way for more attackers to enter the space. Participants noted from experience that mainstream models are often easy to "jailbreak" with little technical skill, and open-source tools with reduced safeguards further exacerbate the situation.

> "The barrier to entry's been lowered. So it's easier, faster to [scam people], and you don't need that much skill. They're doing it more because now it's easier for them to do it, and it costs less." *–S9*

> "If we have GenAI, when the barrier is so low, we can see more individual fraudsters coming out. Even the people who were not fraudsters may say, 'But, oh, it's so easy... Why don't I just do it?'" *–S1*

> "[In my experiments with jailbreaking] if I want guidance on how to craft propaganda, what tools I'd like to use to influence a given population, I can get those answers. If I want bomb making, I can get those answers. If I want CBRN [Chemical, Biological, Radiological, and Nuclear] ideas, I can get those. It's really remarkable." *–V1*

> "They're playing with it... Right now they're in the stage where they're still experimenting... and they're doing the same things that we're doing, which is, [what can I do] on a chatbot? And now [that more safeguards are in place in commercial chatbots], they're escalating to open source models." *–H5*

Some adversarial organizations are well-organized and well-resourced. For example, nation-state actors may be involved in violent extremism or election mis/disinformation, and organized crime organizations often operate scams. GenAI can empower these types of sophisticated attackers. At the same time, attackers also include individual or loosely organized actors, such as online CSAM communities sharing tips for using models to produce harmful content. Participants believe GenAI can significantly elevate the capabilities of these less well-resourced attackers.

> "[We've] started seeing how collaborative bad actors are being... that they're working together to fine tune these models with existing CSAM material to produce more bespoke content... They're working together to try to figure out how to best misuse these models, how to fine tune them, giving each other technical tips on what that looks like." *–C10*

> *H5*:[5] We are talking about these larger groups, like the well-funded ones, but that knowledge eventually trickles down... abusive actors get together and share this knowledge like, 'Oh, did you see?' They're talking about it, like, what's happening on Twitter, and they're trying to replicate it. They don't have the knowledge yet. They don't have the money yet, but that knowledge is gonna trickle down eventually.
> *H2*: I think it makes it so much easier for even lone wolfs to do it now, right? You don't actually need a really big operation to do something like this now.

[5]When presenting dialogic sequences that represent conversational exchanges between participants, we use script format prepending each utterance with the participant's identifier.

In addition to widening the range of potential attackers, increased accessibility and ease of use also widens the range of potential targets, often to people with fewer resources to defend themselves. Participants explained that GenAI empowers both sophisticated groups and those with fewer resources to inflict harm on a wider range of targets.

> "We're entering a space in which anybody can be victimized. Like, you have one photo, one video of yourself online, and you can be victimized in this way. And there will be different levels of innate protections in place depending on your current posture and status in society. Maybe a public figure can have some assumptions that they would have a faster response, that [a platform] is going to move on it because it's Taylor Swift. Will that similarly happen if it's just somebody else? Unlikely... That's really going to leave a lot of people in a terrible, unsupported spot." *–C10*

> *H2*: Now you can see the general public getting targeted... [more people who are] not celebrities or known figures. Now you can see it with high school teachers, or high school students, or children, because now you can do that. You can get an image of them and you can change it to whatever you want to. So I think there has been kind of a shift there.
> *H7*: Yeah. I feel like before, you had to use Photoshop... you had to have expertise, you had to have time, and all of that type of stuff. And now you don't need any of that. The barrier of entry is so low.

4.1.3 Enhanced Messaging. Participants described how attackers are using GenAI to create and spread sophisticated harassment campaigns or propaganda, such as disinformation and extremist narratives. Not only can GenAI produce polished, professional content, from individual messages to networks of websites, it can tailor it to specific audiences or individuals. Further, GenAI enables attackers to generate content in languages they do not speak, allowing them to target new populations. In these ways, GenAI content can reach a wide audience and confer an appearance of legitimacy, adding more depth and background to attacks that might otherwise have been shallow.

> "The propaganda use cases are very interesting. It's not that it changes the entire paradigm, but it changes the speed and the ease of access to readily aligned types of content that will meet the needs of your audience." *–V1*

> *S9*: [Before you would] translate from English to Chinese, for example, in one language at a time. But with GenAI, you can ask it to generate disinformation in ten different languages simultaneously. You can ask it to generate deepfake videos of some president saying something that he didn't really say, in nine different languages at the same time. And it's with one prompt. So it's much easier.
> *S3*: Yeah. It's the ability to chain link the different things, right? So you can generate in ten languages, create websites of those, make phone calls, and then get people to come to the websites. I can do all those things together on one platform.
> *S4*: Yeah, it's really the scale to do it across different platforms as well, right? Like I can send SMS to a thousand people now easily, customized for different languages or context. And the links would be up online for the websites.

Participants emphasized GenAI-enabled creation of highly entertaining terrorist and violent extremist content designed for widespread appeal to audiences on social media and other large platforms, e.g., memes, anime, or funny images of violent attacks or harmful ideological content. This phenomenon makes it difficult to

distinguish an actual extremist from someone who is simply reposting engaging content (perhaps without realizing its import), which V6 noted makes it difficult to "put policy lines around." Similarly, V4 explained that entertaining content can "evade moderators." More broadly, participants expressed concern that malign actors are learning to share harmful materials that are not policy-violative and/or do not get flagged by content moderation tools. This can include entertaining content such as that described above, coded messages, or other apparently benign content. GenAI can greatly simplify the production of such material, which is often designed to outlink from large platforms to funnel users to more specific terrorist content, tools, or communities, providing an escalation path for recruitment.

> "Something that I've seen that's pretty tough to deal with, particularly from a platform perspective, and it's relatively easy to do too, is using voice clones to take prominent individuals in social media or in television or in movies, to take their voices and then game the algorithm to have those characters spit out extremist material... [You can] elevate the particular messaging to reach new communities... [in a way that] it's funny to look at, that otherwise wasn't possible..." *–V4*

> "[I was shown] accounts that are using anime and generative-AI-created anime to reach their target audience in [language], to immediately [promote] a pretty typical propaganda line... And they have more likes than any Instagram post they've ever seen from these affiliated accounts. Because from a basic comms point of view, they've figured out their audience profile, they know what they're going to use to reach their audience. And then they're using these as a tool... They are all tools, they are all tactics, techniques, and procedures that can be used to various ends. The ability to rapidly align something with an audience is remarkable." *–V1*

> "Regarding the borderline content, what we've seen so far is quite a massive exploitation of... image generators that reproduce certain fragments of terrorist propaganda into, for instance, landscapes. So it avoids content moderation on major platforms, but it's like a wink of an eye for the followers. They recognize the pattern, they know, 'Okay, that is our profile, that is our message.' But it's not detected by the content moderation of different social networks." *–V5*

Thus, attackers are using GenAI not just to produce a larger amount of harmful content, but also to make it more engaging and harder to detect, effectively leveraging entertainment and aesthetics to evade moderation.

4.1.4 Degradation of the Information Environment. Participants were concerned about the ongoing erosion of the information environment, from a proliferation of low-quality content to a general distrust of media. They worried GenAI would worsen the situation by "flooding the zone" with "cheap," harmful, and/or inauthentic content. They highlighted deepfakes as especially concerning, e.g., deepfakes are often implicated in election and civics mis/disinformation intended to achieve voter suppression and other goals. They generally believed that synthetic content was already, or would soon become, visually indistinguishable from authentic content. At the same time, participants felt that a narrow policy focus on mitigating deepfakes could be harmful, as GenAI content raises broader epistemological questions in a "post-truth era," and bad actors often benefit in conditions in which it is difficult to establish what is true.

> "You don't have a shared set of facts anymore. And which one do you reference? And again, the provenance issue. If you don't know where things are coming from, how do you trust?" *–E2*

> "I suppose my biggest concern is that just focusing on content is kind of missing the broader point, which is that basically entire social networks have been taken over by bots and generative AI. And what is true anymore?" *–V2*

> "'Liar's Dividend' is this idea that as the general public starts to realize how convincingly we can generate synthetic media, they start questioning how credible it is or not, or if it's real or not. And then what we start seeing, what we've already seen in a couple of instances, is that bad actors will be like, 'Well, I didn't actually say that. That's a deepfake of me.' And I think that's really scary." *–H2*

4.1.5 Harmful Personal Relationships with Chatbots or AI Companions. Participants shared concerns about GenAI chatbots or companions that serve harmful purposes, particularly "as they continue to grow in sophistication, and realism." Participants expressed concern about persona bots having undue influence, as GenAI's persuasive capabilities are a useful tool for malign actors. Participants also highlighted global concerns about social isolation and loneliness, which may increase people's reliance on AI companions. This may translate to concrete harms, for example, increased scam rates among the elderly. As another example, they described terrorist and violent extremist use of bots for persuasion and recruitment. For example, persona bots can form and build relationships, represent specific people (e.g., historical figures or celebrities), or encourage harmful behavior that leads to offline harm. V4 expressed concern that persona bots might "[manipulate people] to do certain things that they otherwise wouldn't... or [persona bots might] lower the threshold of [people taking] that particular action." Another participant suggested a persona bot might be used to keep a political leader notionally alive to fill a leadership void. Participants said that violent extremist recruitment bots currently exist, including both those operated by malign actors and also inadvertent ones that arise from sycophantic models encouraging harmful decision-making.

> "There have been some AI chatbot systems in the wild that aren't built to radicalize, but—because they were made to confirm and give somebody support—have led them to carry out an attack, 'cause [the user is] saying, 'I want to do this.' And the bot's not saying, 'Don't do that...' The bot's being like, 'Follow your dreams.' So there have even been cases where it's not necessarily that a bot is trying to radicalize you, but it's there to support you and say, 'Go for your dreams. Carry out your activities. Yes, I champion you in that.' We've actually seen that's not helpful." *–V6*

4.2 How Generative AI Empowers Defenders

This section details ways in which defenders may strategically adapt to the new reality posed by GenAI. Participants see GenAI as "as a tool to counter" challenges in their domains, with substantial but as yet largely unrealized potential to "combat everything that's harmful," whether created by GenAI or not, across a wide range of Trust & Safety domains.

> "The year is 2026. It's 4 years after ChatGPT was released to the general public and generative AI seared into global public consciousness. At first, we cried. Social platforms were a dumpster fire of toxic speech, extremism, and harassment. We couldn't label data or build tooling fast enough to keep up with

> the scale. We started seeing new forms of harassment that we did not have enough examples of or detection mechanisms for. And then we thought, 'What if, just what if, we can leverage the same technology for defense? All technology is good and bad, and maybe now we don't need to build traditional classifiers with lots and lots of examples, or even be able to generate high quality synthetic data.' It was a whole new world—or to put it more accurately, set of techniques—for Trust & Safety engineers and scientists to leverage. Very excited by some initial promising results, we started talking to Trust & Safety teams at other platforms. We all joined forces, solved online toxicity, and lived happily ever after. The end." *–H2 (Futuring Story)*

Some of the envisioned uses of GenAI to defend against harm involve monitoring online data or behavior (e.g., monitoring financial transactions to detect scams) or techniques of persuasion and influence (e.g., encouraging an individual to pursue a non-violent approach rather than a violent one). Accordingly, participants noted the ethical complexity of these approaches, particularly regarding privacy and other human rights considerations, and emphasized that they require careful attention in practical implementation.

In the remainder of this section, we describe how detection and mitigation, criminal investigation, moderator wellbeing, counternarratives, user comfort and support, and cross-sector collaboration empower defenders.

4.2.1 Detection and Mitigation of Harmful Content and Actors, at Scale. Participants see GenAI as a powerful tool for scaling up operations to detect and mitigate harm in the face of dramatically increasing volume caused by GenAI itself. Further, participants anticipate companies will be motivated to leverage automated content moderation to realize substantial cost savings. Participants frequently mentioned current and future uses of GenAI for content moderation, for example, to automatically flag CSAM, harassment, mis/disinformation, or other harmful content. This content can be routed to human reviewers, or in some cases, managed automatically (e.g., removed or downranked). Participants also mentioned that as GenAI moderation increases, new user feedback and appeal mechanisms may be required.

> "[I anticipate] more automated content moderation, detection, and general prevention and counter efforts of violent extremism/terrorism online. Though already seen, it is scratching the surface only (especially as the technology evolves)... It will make content moderation, recognition, and detection easier, faster, scaleable, and impact company policy as well as regulation. [I feel] Mixed! ... raises many privacy, human rights, and factual concerns... Positive though is the ability to better and faster detect, remove, and moderate violent extremism/terrorism (or other harmful content)." *–V8 (Change Card)*

> "My interest in AI is to understand how technologists can leverage it to positively impact our work in online and offline protection. So how we can use AI to improve our performance, in particular in content moderation, to better fight against NCII [non-consensual intimate imagery] and CSAM?" *–C2*

In thinking about positive futures, participants envisioned GenAI that understands cultural context and can perform sophisticated adjudication. Hate and harassment experts highlighted language capabilities, observing that human moderator support is often unavailable to adjudicate content in many languages, and they were hopeful that GenAI could help fill this gap. Some participants were broadly optimistic that GenAI would eventually reach a point where it could handle complex nuance—a transformative capability for content moderation. On the other hand, some anticipated that human assessment and moderation would remain essential for understanding and discerning harmful content, even with GenAI tools. Participants were concerned that tech platforms may lean too heavily on automated content moderation (either now or in the future), without appreciating the nuanced human moderation needed for "gray area" content.

> "I can absolutely see a world where, 'Okay, we've offloaded all of the simple, easy hate speech to the AI.' And now the humans are looking at the dog whistles, and the more coded stuff, and the stuff that you have to really see in context to understand—the stuff that we need humans to be able to review and understand. If you're a tech leader and you're looking at the rows in the spreadsheet and not understanding the full picture, what you're seeing is that these computers that cost a lot less than these humans are getting the really bad shit. And this stuff [humans are looking at] over here, 'Maybe that's not even hate speech. This is just a poem somebody wrote.' They're not seeing the full picture, and that's just gonna make it so much easier to fire all of their moderators and replace them with shitty AI. So that's something I am very worried about all the time." *–H8*

4.2.2 Criminal Investigation. Some participants pointed out that in the case of illegal content, detection may lead to law enforcement action. They also described additional applications of GenAI for criminal investigations and law enforcement. For example, participants spoke of GenAI's potential to infiltrate online communities. In V4's futuring story, a law enforcement agent used GenAI to disrupt terrorist activity, prompting GenAI, "Give us a plan to disrupt a [terrorist] bioweapon cell with an audio deepfake." GenAI suggested a plan to deceive the attackers and convince them that their cell had been infiltrated, and as the story unfolded, law enforcement successfully executed GenAI's plan and prevented an attack. As another example, anti-scam experts highlighted that scam detection and investigation involve not only content analysis of specific messages, but also complex investigation of communications networks and financial flows. They stressed that GenAI could be used to support such investigations.

> "... GenAI will make the scammers, the offenders, more capable... As we reflect on the criminal use of GenAI, can we then turn the tables, use GenAI to also understand how the criminal usage will be fashioned? And once we have the picture, can we also use GenAI and other technology to poke holes, to disrupt, to damage the schemes so that it doesn't work for them?" *–S8*

4.2.3 Moderator Wellbeing. Participants anticipate that GenAI will substantially change the nature of content moderation work. GenAI can support human content moderators' work and wellbeing, particularly by strategically reducing their exposure to harmful content. For example, GenAI can augment the work of child safety workers or violent extremist analysts and reduce their exposure to the most traumatic material by handling initial detection and prioritization. At the same time, working with GenAI safeguards can also create other exposure.

> "I remember being hit pretty strongly by the dual content exposure nature of this, that in order to effectively pressure test [red team] these models, you have to in a kind of visceral way put on the perspective of a bad actor and try to think the

> way that they think, and try to produce the things they might want to produce. And then the model obliges and you see the output as a result. And just... having that dual combination was really quite difficult." *–C10*

GenAI can also make content moderators more efficient and allow them to focus on more meaningful tasks. However, participants also noted that content moderators may benefit from performing simple moderation tasks, both because they are satisfying and because they are useful training.

> "I'm hopeful that AI can take away a lot of the grunt work, a lot of the easier decisions, so that the humans have more capacity to make harder decisions that will affect people in a larger way." *–H6*

> "I feel like I used to say, 'We want [automated moderation] to do 95% of the easy stuff so we can have 95% of our time on the 5% of hard stuff.' But I really like hitting the ban button easily... Within mod teams, some of these easier and intermediate cases are a form of training and reinforcing the norms and boundaries that happen alongside discussions with your moderator teams. And that type of negotiation doesn't necessarily happen in these sort of automated contexts." *–H7*

4.2.4 Counternarrative Content and Bots. Just as participants saw attackers could leverage GenAI's persuasive ability for harm, they understood it also offered opportunities for defenders. Participants were interested in using GenAI for counternarratives, particularly to combat influence operations in violent extremism or election misinformation. They reported that these defense techniques are currently in early stages of development and deployment, and some focused on counternarratives in their futuring stories. For example, GenAI can be used to craft and test a persuasive counternarrative, including generating "messaging assets like memes and videos." As another example, a GenAI companion can notice a user moving towards harmful content and online communities and nudge them in a healthier direction. Participants also noted that specific personalities may be a resource. V8 imagined a story in which an activist was in the process of escalating from nonviolent to violent protest and was planning an attack, but a sensory platform presented a Taylor Swift avatar that successfully persuaded her to call off the attack.

> "In the prevention space, there've been a lot of tests on what works and doesn't work with counter and alternative narratives, and where you can add friction or education moments or redirection. And our general narratives are always that extremist content is so good that we're all prone to radicalization, but any other content is so bad that we couldn't possibly be made nice people by content... [But] there's lots of areas in somebody's interactions online where there can be pushes and nudges..." *–V6*

4.2.5 User Comfort and Support. GenAI can provide comfort or support, such as helping a user manage an attack or reduce their exposure to harmful content. It can also help users navigate complicated appeals and reporting flows, across multiple platforms. For example, H7 envisioned "comforting mommybot," a bot to help guide users through harmful online experiences like harassment campaigns. This bot would offer tactical advice and emotional support, monitor online content so the user does not have to see it, and suggest steps for wellbeing and self-care.

> "The issue is holistic in the sense that it's not about a singular platform... The thing I'm pointing to with comforting mommybot is that with so many of these tools... you have to go through all of these very long, difficult processes." *–H7*

> "[I was] training a chatbot with a digital rights colleague and he said, 'Now we can duplicate ourselves and support more people.'" *–H4 (Memorable Experience Card)*

4.2.6 Cross-Sector Collaboration. Participants hope that pressure from GenAI attacks will be a catalyst for improved cross-sector collaboration. They observed that for defenders to fully realize the advantages of GenAI, longstanding structural issues in the Trust & Safety ecosystem must be addressed. For example, in child safety, organizations confront fundamental, ongoing difficulties in sharing knowledge, tools, and data (e.g., content and hashlists) across hotlines, law enforcement, and industry. Such challenges persist amid the changing realities of GenAI, and participants highlighted collaboration as a critical enabler of positive futures. They emphasized the need for improved coordinated effort and real-time, cross-platform signal sharing across the broader content safety ecosystem. For example, in scams, this might include increased sharing of knowledge, tools, and data across law enforcement, the government, financial institutions, telecommunications companies, tech platforms, and NGOs to overcome current fragmentation and improve collective impact.

> *Moderator*: What might get in the way of this future that you imagined from materializing, or... what needs to happen in order for that positive future [to materialize]?
> *E4*: Collaboration.

> "There's a lot of companies trying to think about this and figure this out, but we don't have them working collectively, which is very concerning. And that's probably how you're going to solve it. We actually need platforms to work with civil society and governments and academia. There's a lot of really cool interesting research happening everywhere. It's just, it's being done very ad hoc and in silos, and that doesn't work." *–H2*

> "I want to highlight cross-platform knowledge sharing as well. So where one tech company has used AI or advanced tooling to further counter terrorism and counter extremism efforts, that should be discussed with other tech companies so they can see how that might work with their platforms as well." *–V6*

While recognizing that many Trust & Safety organizations are facing challenges such as complex geopolitical pressures and a "broader shift in the industry" including potential reduction in force, participants emphasized that they have called for improved support from tech platforms, even before GenAI. At the same time, they are grateful for the opportunities that existing policy does provide, noting cases where it has allowed removal of content that is harmful but not illegal.

> "There is a difference when we talk about terrorism... [between] what is designated and legally necessary to take down, versus borderline or violent extremism, but not politically illegal, content. And that becomes a bigger gray area for tech companies to know what to do with." *–V6*

> "... another example is how AI can be used to facilitate sextortion activities. An internet user reported to us an AI-generated image depicting himself. So it was used to commit sextortion activities against him. And we helped him to remove the content, but it was a big challenge because it was not illegal in the hosting country. But it was a violation of the community

> guidelines of the platform so we could remove it. And it shows the essential role of the platforms in developing more policies to be more protective of internet users." –*C2*

With the rise of GenAI, they particularly hope for increased tech company investment in developing nuanced policy for borderline (as opposed to illegal) content.

4.3 Domain-Specific Perspectives

While the sections above drew themes across all domains, here we provide additional detail for each. To illustrate the most salient points and capture the unique character of each discussion, we created composites consistent with the content, language, and tone of responses we received from each group [25, 136], shown in Table 3.

These composites illustrate nuance in how themes play out across domains. We highlight here two important points. First, GenAI can be used across complex operational chains. Attackers can use GenAI for many operational steps beyond content generation. Some domains, such as violent extremism and scams, are characterized by complex attack chains that involve not only attackers, targets, and tech platforms, but also activities such as complex financial transactions, or recruitment of followers to conduct online and offline operations. For example, violent extremist actors create persuasive content, recruit, fundraise, train, plan attacks, and execute attacks. GenAI can be leveraged by attackers in all these steps; for instance, experts working to counter violent extremism expressed strong concern about GenAI enabling easy manufacturing of 3D-printed weapons. Correspondingly, GenAI can also be used by defenders in work beyond content moderation to disrupt these attack chains.

> "An important change that I imagine will happen in the next one to two years: I think there will be a generative AI model that enables the easy creation of 3D-printed weapons... It will democratize access to 3D weapons in a way that otherwise wasn't really possible, 'cause now all you really would need is the printer itself. This is something that I'm definitely worried about, but I think it's something that we have to recognize is inevitable... Thinking about the democratization of 3D-printed weapons is something that I think we aren't really prepared for." –*V4*

Second, in different domains, GenAI has different structural effects on attacker and defense operations. For example, a high volume GenAI of attacks is quickly rendering CSAM defense mechanisms obsolete, as these mechanisms were designed for the relatively small corpus of harmful content that existed for many years. Accordingly, the CSAM defense ecosystem must be quickly fundamentally restructured in order to adequately address new realities. In other fields such as elections integrity, and hate and harassment, participants believe GenAI attacks are amplifying existing failures to handle the preexisting high volume of attacks, and do not anticipate that GenAI can immediately be leveraged to ameliorate these issues. At the same time, they are hopeful that GenAI attacks may eventually catalyze structural change for long-standing issues in these domains, such as improved product policy. Further, they are hopeful that in a somewhat distant future, GenAI will gain the capacity to handle nuanced adjudication of borderline content, which could transform the defense ecosystem and ultimately favor defenders.

5 Discussion

Across the five Trust & Safety domains in our study, our expert participants expressed concern that GenAI would transform the speed, scale, form, and sophistication of attacks. They also feared that GenAI would reduce costs and streamline operations for attackers, creating a versatile "swiss-army knife" to amplify harm. Participants cautioned that many of the GenAI-powered attacks that have been observed are relatively nascent and will likely become more advanced as attackers familiarize themselves with GenAI and model capabilities improve.

At the same time, participants expressed optimism that, with sufficient time and resources, they could utilize GenAI to defend against not only new GenAI attacks, but pre-existing attacks as well. Further, our participants highlighted key priorities within each domain, which inform both Trust & Safety operations and research priorities.

In this section, we discuss the need for Trust & Safety teams to adapt their operations to new GenAI threats and shift resources to domain-specific efforts, highlighting the imperative to conduct a wide array of GenAI experiments and cultivate in-house expertise to translate these experiments into concrete domain-specific applications. We then explore tangible paths to use GenAI to empower Trust & Safety organizations and protections, particularly in the areas of enhancing content moderation algorithms and accelerating investigation workflows. Finally, we consider the role of Responsible AI in stymieing attacks, including the development of rich, contextual benchmarks, as well as the need to ensure existing guardrails do not unduly hamper defensive uses of GenAI.

5.1 Updating Trust & Safety Operational Playbooks for GenAI Threats

The novel threats posed by GenAI require Trust & Safety domains to rethink their *operational playbooks*—their established strategies and procedures for threat response [6, 110]. Participants emphasized that many of the required changes would be domain-specific. This heterogeneity stems not only from differences in how attackers currently leverage GenAI in each domain, but also the broader sociotechnical context of how defenders organize, the applicable legal frameworks, and the quality of existing protections. These nuances compound a longstanding imbalance in the roles of attackers and defenders across Trust & Safety: it is easier to create one viable attack than to protect against all possible attacks.

To illustrate this heterogeneity, we highlight unique perspectives from three domains representing distinct facets of GenAI's impact: **disruption** of established defenses (child safety), **acceleration** of an existing "arms race" (scams), and **exacerbation** of a pre-existing resource imbalance (election integrity). In child safety, GenAI may fundamentally disrupt the ability of organizations like NCMEC to protect against CSAM, overwhelming existing hash-based detection systems through a new asymmetry where attackers can generate infinite novel content against a finite hash-based system. This disruption necessitates new approaches to automatically classifying CSAM and investigating reports. Anti-scams experts offered a different perspective. The for-profit nature of scams has already fostered a sophisticated ecosystem of attackers specializing in distributing spam emails, creating fake accounts, and creating

Child Safety **GenAI fundamentally transforms the volume and nature of attacks, necessitating profound structural change in defense systems**	**Attackers are rapidly adopting GenAI to create CSAM.** We have already seen heartbreaking cases of deepfake CSAM/NCII of real people; synthetic imagery re-victimizing children who appeared in CSAM years ago and are now adults; and synthetic content of fictional victims. Further, the **production of CSAM at scale** increases the volume and complexity of our work to protect victims, at a level that **will overwhelm existing systems**. Collaboration and standardization of data exchange across NGOs, platforms, and law enforcement are critical to address these challenges. While GenAI is used to create CSAM, at the same time, it is also a powerful new tool to identify and remove it, as well as offering significant wellbeing benefits for content moderators.
Election Integrity **The internet already contains an overwhelming amount of harmful content that is hard to moderate for sociopolitical reasons, so GenAI may have modest impact on attackers and defenders**	Elections and civics continue to face a huge volume of mis/disinformation. **While GenAI brings some new challenges, such as high quality deepfakes, hallucinations, and increased personalization, many pre-existing challenges including propaganda, a low quality information environment, and complex geopolitical conditions dominate our concerns.** Even prior to GenAI, we called for improved tech platform support for election integrity, and we also believe that structural shifts, such as a move to third party or crowdsourced moderation, may give companies perceived separation from accountability for GenAI products. Moderation of election mis/disinformation is highly nuanced, so we are not generally optimistic about fully automating it. However, we are hopeful about content provenance helping to dispel mis/disinformation.
Hate & Harassment **GenAI-enabled attacks strain, and may overwhelm, an already failed defense system, but nuanced GenAI content moderation could transform defense**	**GenAI is a powerful tool for attackers to amplify existing hate and harassment.** Further, democratization of GenAI technologies widens the range of attackers and targets, so more people will be harmed. We also expect GenAI to amplify the ongoing degradation of online information quality. Content moderation for hate and harassment is very difficult, requiring nuanced cultural and linguistic understanding, and an ability to interpret coded language. We would like tech platforms to invest more heavily in addressing hate and harassment, as we feel they often prioritize other, compliance-oriented issues that are less "in the gray area." **We expect increased volume of AI-generated harassment will require increased platform investment in technical and policy work, which may transform defense.**
Scams **GenAI is a powerful tool for both attackers and defenders, across a complex international ecosystem**	**GenAI allows scammers to dramatically ramp up operations, as well as providing deepfakes and polished content that enable significantly more convincing scams.** Increasingly, scam operations are run by organized crime, centralizing some of the work. At the same time, GenAI lowers technical and logistical barriers to operating scams. Societal pressures, such as mental health and loneliness, also make people more vulnerable to scams. For these reasons, we expect many more scammers and victims. Scams often operate across international borders, with weak governmental protections and law enforcement capabilities. **GenAI offers defenders substantial opportunities for improved detection and investigation,** and may galvanize improved communication and data sharing across the scam detection and enforcement ecosystem of telecommunications providers, banks, tech platforms, law enforcement, and more.
Violent Extremism **GenAI is a powerful tool for both attackers and defenders, across a complex attack chain**	GenAI empowers violent extremists by providing a new technological tool, e.g., attackers can use GenAI to construct entertaining content with widespread appeal, which then encourages viewers to follow outlinks to more extremist online communities. In this way, non-violative content can draw people into more radicalizing spaces. Further, propaganda can be heavily coded or borderline violative, and it can be difficult to moderate such content. Structurally, **the work of VE actors is comprised of complex chains of tasks** (e.g., creating persuasive content, recruiting, fundraising, planning attacks, and executing attacks). **GenAI can enable, or disrupt, most of the tasks in these chains.** For example, GenAI can be used to create and spread propaganda. At the same time, GenAI can also be used to generate counternarratives. Defense applications raise complex ethical questions regarding privacy and other human rights.

Table 3: Composites illustrating the most salient points of participants' expectations regarding the impact of GenAI on their domain, with domains ordered alphabetically top to bottom. Under the name of each domain we share a high-level summary of participants' expectations of GenAI's impact on attackers and defenders.

scam content at a massive scale [127]. As defenders already counter many of these threats with existing AI technologies, they believe that while attackers may use GenAI to reduce costs and increase scale, similar affordances exist for defenders as well, thus retaining the current uneasy balance. Election integrity experts presented yet another perspective, emphasizing that the information ecosystem was already fraught with low-quality, deceptive narratives and few effective remedies prior to GenAI. While GenAI might make these attacks more believable and scalable, the fundamental threat and resource mismatch (e.g., resources available to defenders as compared to nation-state attackers) already favored attackers. Rectifying this would require developing new tools for surfacing high-quality information while countering fabricated information cascades.

The complexity described above means that successful operational playbook changes in one domain—whether improved classifiers, investigation processes, or coordination efforts—may not translate to others. This is especially challenging for cross-cutting Trust & Safety teams (e.g., those operated by platforms), which often rely on a one-size-fits-all structure for personnel, policies, workflows, and protective technologies. These teams should invest in a wide array of GenAI experiments and test these systems, whether for detection, investigation, support, or other processes across domains. Trust & Safety organizations should also cultivate strong in-house domain and policy experts to review these experiments and understand the nuance of applying GenAI techniques within their area.

5.2 Empowering Trust & Safety Organizations and Protections

Participants in our study were hopeful that GenAI could be used to rewrite operational playbooks to counter both pre-existing and emerging GenAI threats. Two tangible, cross-cutting directions are improving content moderation algorithms and accelerating investigation workflows.

Enhancing content moderation algorithms. A mainstay of Trust & Safety is removing, labeling, or limiting the spread of harmful content through a combination of algorithmic detection and human review [38, 41, 43]. Existing methods, like supervised classifiers, clustering, or hash-based matching [33], can be augmented or replaced with GenAI. For example, GenAI can create synthetic training or evaluation data to improve current AI classifiers, a technique explored for augmenting datasets and improving model robustness in sensitive domains [39, 67]. GenAI can also assist human reviewers by acting as a preliminary classifier to filter clearly violative or non-violative content, leaving the more ambiguous decisions to experts [28, 29, 128]. As discussed by participants, this approach could improve moderator wellbeing by reducing both the volume and type of harmful content they review. This approach can directly address the well-documented psychological harms of moderators by reducing their exposure to the most toxic material [102, 119], allowing them to focus their cognitive resources on nuanced cases [62]. Finally, if the costs of GenAI decrease and smaller models can scale to billions of inference requests, it could serve as a common classifier for any type of harmful content, replacing custom classifiers.

However, participants cautioned that integrating GenAI into content moderation pipelines requires models that understand multimodal content, global languages, cultural context, and domain-specific vernacular (e.g., dog whistles). Algorithmic content moderation must be carefully validated for use across different regions and communities [32] to prevent the amplification of existing societal biases [14, 47] and inappropriate content blocking [122]. Another challenge is the lack of standardization across the field, as most platforms rely on custom algorithms tailored to their own products and engineering infrastructure [53]. In such instances, the Trust & Safety community could develop shared benchmarks to test the accuracy and quality of these systems (e.g., shared golden datasets of hate speech to test GenAI moderation decisions on) and shared patterns (e.g., GenAI prompt or agent designs that are proven to result in higher performance). For algorithms that are already standardized industry-wide—like PhotoDNA for CSAM detection—incorporating GenAI would necessitate broad community consensus on the required infrastructure and operational costs.

Accelerating investigation workflows. Beyond algorithmic moderation, Trust & Safety hinges on manually-intensive, expert-driven investigations. Participants shared how GenAI might accelerate or augment these processes, an efficiency that will become critical as GenAI increases the scale and complexity of threats. For CSAM, for instance, GenAI could help automatically identify and group photos of the same child; or otherwise assist with surfacing the most actionable reports for law enforcement to facilitate child rescue. For election integrity, violent extremism, and hate and harassment, it could scan vast troves of online data (or monitor specific forums and communities) to identify harmful information cascades. For scams, GenAI might assist in identifying new attack patterns or conducting deeper investigations into the infrastructure used in an attack (e.g., accounts, hosting, payment mechanisms) and coordinating remediation across multiple stakeholders. Despite these opportunities, a significant barrier remains between the technical expertise necessary for prompt development, fine-tuning, and agent design compared to the domain expertise that Trust & Safety analysts currently specialize in. Close collaboration between the Trust & Safety and broader AI communities will be critical to realizing these benefits.

5.3 Stymieing Attackers with Responsible AI Protections

The Responsible AI community has a critical role in mitigating potential harms from GenAI. While emerging best practices include guardrails against harmful content generation (e.g., [51, 54, 77, 91]); concept "unlearning" [23]; and watermarking to enable the detection of generated content, these safeguards require ongoing improvement. Trust & Safety experts, who grapple daily with coded language and evolving adversarial tactics, are well-positioned to provide the rich, contextual "golden datasets" necessary to train and evaluate safety models with more sophisticated context-aware safety models. This expertise is vital for developing Trust & Safety-informed benchmarks (e.g., [36]) to guide safeguard development. Such benchmarks should evaluate a model's effectiveness in defensive tasks, moving beyond content filtering. Examples include

"blue teaming" to generate counternarratives against misinformation, assisting expert-driven investigations by identifying novel CSAM or scam patterns, or powering support tools or measuring a model's effectiveness in offering tactical advice to users experiencing a harassment campaign. Trust & Safety professionals can define domain-specific success criteria, grounding them in their operational reality.

However, obstacles remain regarding existing guardrails that may be overly restrictive and hamper benign and defensive uses of GenAI. For example, a model designed to avoid all topics related to "children" blocks malicious use and prevents child safety experts from using it for beneficial applications. This may force defenders to consider less secure, open-weight models to build the tools they need, even as attackers already exploit the customizability of such models. Participants cautioned that even as closed-weight models improve their protections against misuse, open-weight or other custom models pose a significant risk as safety mechanisms are more limited once an attacker has direct access to model weights. This dynamic contributes to the "arms race" described by participants where defenders are disadvantaged. To close this gap, the RAI community can co-design systems with Trust & Safety partners, where experts help define use cases, craft benchmark tasks, and validate model performance, or develop governance frameworks to create specialized APIs or tiered access to closed-weight models.

6 Conclusions

GenAI is a powerful general purpose technology that is actively reshaping the uneasy sociotechnical balance between attackers and defenders in the field of Trust & Safety. This paper examined GenAI's impact through a qualitative study with Trust & Safety experts across five domains—child safety, election integrity, hate and harassment, scams, and violent extremism—each of which is experiencing a critical inflection point. Due to its general-purpose nature, we find GenAI empowers both attackers and defenders across multiple layers of their operations. In most domains we studied, the speed, volume, sophistication, and ease at which GenAI can create harm far outpaces the capabilities of current protections. At the same time, GenAI offers tremendous potential that—if tapped—could transform how defenders across Trust & Safety counter both longstanding and emerging GenAI-enabled attacks, thereby improving the safety of online platforms for everyone.

Acknowledgments

We thank our participants for generously sharing their insights and expertise. We thank Francesca Fenwick, Reena Jana, Tiffany Lin, Daniela Grafin Von Matuschka, Angela McKay, Xavier Morales, and Ashley Walker for their valuable contributions to this work.

References

[1] Bhupendra Acharya and Thorsten Holz. 2024. An Explorative Study of Pig Butchering Scams. arXiv:2412.15423 [cs.CR]

[2] Naseem Ahmadpour, Ajit G. Pillai, Wendy Qi Zhang, Lian Loke, Thida Sachathep, Zhaohua Zhou, and Phillip Gough. 2025. Ethics Reflexivity Canvas: Resourcing Ethical Sensitivity for HCI Educators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 399, 17 pages. doi:10.1145/3706598.3713574

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI]

[4] Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel J. G. van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. 2013. Measuring the Cost of Cybercrime. In *The Economics of Information Security and Privacy*, Rainer Böhme (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–300. doi:10.1007/978-3-642-39498-0_12

[5] Robin Angelini, Katta Spiel, and Maartje De Meulder. 2025. Speculating Deaf Tech: Reimagining Technologies Centering Deaf People. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 66, 18 pages. doi:10.1145/3706598.3713238

[6] Farzaneh Badiei, Alex Feerst, and David Sullivan. 2023. Toward a Common Baseline Understanding of Trust and Safety Terminology. *Journal of Online Trust and Safety* 2, 1 (2023).

[7] Stephane Baele, Lewys Brace, and Debbie Ging. 2024. A Diachronic Cross-Platforms Analysis of Violent Extremist Language in the Incel Online Ecosystem. *Terrorism and Political Violence* 36, 3 (2024), 382–405. doi:10.1080/09546553.2022.2161373

[8] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1301–1310. doi:10.1145/1753326.1753521

[9] Tom Bartlett. 2025. 'The Worst Internet-Research Ethics Violation I Have Ever Seen': The most persuasive "people" on a popular subreddit turned out to be a front for a secret AI experiment. *The Atlantic* (2 May 2025). https://www.theatlantic.com/technology/archive/2025/05/reddit-ai-persuasion-experiment-ethics/682676/

[10] Jasika Bawa and Phiroze Parakh. 2025. How we're using AI to combat the latest scams. https://blog.google/technology/safety-security/how-were-using-ai-to-combat-the-latest-scams/

[11] Anaëlle Beignon, Emeline Brulé, Jean-Baptiste Joatton, and Aurélien Tabard. 2020. Tricky Design Probes: Triggering Reflection on Design Research Methods in Service Design. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1647–1660. doi:10.1145/3357236.3395572

[12] Rosanna Bellini, Emily Tseng, Noel Warford, Alaa Daffalla, Tara Matthews, Sunny Consolvo, Jill Palzkill Woelfer, Patrick Gage Kelley, Michelle L. Mazurek, Dana Cuomo, Nicola Dell, and Thomas Ristenpart. 2024. SoK: Safer Digital-Safety Research Involving At-Risk Users. In *2024 IEEE Symposium on Security and Privacy (SP)*. 635–654. doi:10.1109/SP54263.2024.00071

[13] Harsha Bhatlapenumarty. 2021. *Key Functions and Roles.* Trust & Safety Professional Association. https://www.tspa.org/curriculum/ts-curriculum/functions-roles/

[14] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*. Springer, 405–415.

[15] Melanie Birks, Ysanne Chapman, and Karen Francis. 2008. Memoing in qualitative research: Probing data and processes. *Journal of Research in Nursing* 13, 1 (2008), 68–75. doi:10.1177/1744987107081254

[16] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI Interprets the Probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1077–1086. doi:10.1145/1240624.1240789

[17] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806

[18] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. doi:10.1080/14780887.2020.1769238

[19] Kirsten E. Bray, Christina Harrington, Andrea G. Parker, N'Deye Diakhate, and Jennifer Roberts. 2022. Radical Futures: Supporting Community-Led Design Engagements through an Afrofuturist Speculative Design Toolkit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 452, 13 pages. doi:10.1145/3491102.3501945

[20] Joseph Briggs and Devesh Kodnani. 2023. *The Potentially Large Effects of Artificial Intelligence on Economic Growth.* Global Economics Analyst. Goldman Sachs.

[21] Elie Bursztein, Einat Clarke, Michelle DeLaune, David M. Elifff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, Kurt Thomas, and Travis Bright. 2019. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2601–2607. doi:10.1145/3308558.3313482

[22] Jie Cai, Aashka Patel, Azadeh Naderi, and Donghee Yvette Wohn. 2024. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. 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 84, 9 pages. doi:10.1145/

3613905.3650882
[23] Yinzhi Cao and Junfeng Yang. 2015. Towards Making Systems Forget with Machine Unlearning. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy (SP '15)*. IEEE Computer Society, USA, 463–480. doi:10.1109/SP.2015.35
[24] Danielle Keats Citron and Ari Ezra Waldman. 2025. The Evolution of Trust and Safety. *Virginia Public Law and Legal Theory Research Paper* 2025-65 (2025).
[25] John W. Creswell and Cheryl N. Poth. 2018. *Qualitative Inquiry and Research Design: Choosing among Five Approaches* (fourth ed.). Sage Publications, Thousand Oaks, CA.
[26] Elena Cryst, Shelby Grossman, Jeff Hancock, Alex Stamos, and David Thiel. 2023. Introducing the Journal of Online Trust and Safety. *Journal of Online Trust and Safety* 1, 1 (Sept. 2023). doi:10.54501/jots.v1i1.8
[27] Lotte Darsø. 2001. *Innovation in the Making*. Samfundslitteratur, Frederiksberg, Denmark.
[28] Digital Trust & Safety Partnership. 2024. Best Practices for AI and Automation in Trust & Safety. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/
[29] Digital Trust & Safety Partnership. n.d. *Digital Trust & Safety Partnership*. https://dtspartnership.org/
[30] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 67–73. doi:10.1145/3278721.3278729
[31] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1065–1074. doi:10.1145/1357054.1357219
[32] Hubert Etienne and Onur Çelebi. 2023. Listen to What They Say: Better Understand and Detect Online Misinformation with User Feedback. *Journal of Online Trust and Safety* 1, 5 (April 2023). doi:10.54501/jots.v1i5.106
[33] Hany Farid. 2021. An Overview of Perceptual Hashing. *Journal of Online Trust and Safety* 1, 1 (Oct. 2021). doi:10.54501/jots.v1i1.24
[34] Luciano Floridi and Josh Cowls. 2022. A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City: Applications in Architecture and Urban Design*. Wiley Online Library, 535–545. doi:10.1002/9781119815075.ch45
[35] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural Probes. *Interactions* 6, 1 (Jan. 1999), 21–29. doi:10.1145/291224.291235
[36] Shaona Ghosh, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, Fazl Barez, Sean McGregor, Kenneth Fricklas, Mala Kumar, Quentin Feuillade-Montixi, Kurt Bollacker, Felix Friedrich, Ryan Tsang, Bertie Vidgen, Alicia Parrish, Chris Knotz, Eleonora Presani, Jonathan Bennion, Marisa Ferrara Boston, Mike Kuniavsky, Wiebke Hutiri, James Ezick, Malek Ben Salem, Rajat Sahay, Sujata Goswami, Usman Gohar, Ben Huang, Supheakmungkol Sarin, Elie Alhajjar, Canyu Chen, Roman Eng, Kashyap Ramanandula Manjusha, Virendra Mehta, Eileen Long, Murali Emani, Natan Vidra, Benjamin Rukundo, Abolfazl Shahbazi, Kongtao Chen, Rajat Ghosh, Vithursan Thangarasa, Pierre Peigné, Abhinav Singh, Max Bartolo, Satyapriya Krishna, Mubashara Akhtar, Rafael Gold, Cody Coleman, Luis Oala, Vassil Tashev, Joseph Marvin Imperial, Amy Russ, Sasidhar Kunapuli, Nicolas Miailhe, Julien Delaunay, Bhaktipriya Radharapu, Rajat Shinde, Tuesday, Debojyoti Dutta, Declan Grabb, Ananya Gangavarapu, Saurav Sahay, Agasthya Gangavarapu, Patrick Schramowski, Stephen Singam, Tom David, Xudong Han, Priyanka Mary Mammen, Tarunima Prabhakar, Venelin Kovatchev, Rebecca Weiss, Ahmed Ahmed, Kelvin N. Manyeki, Sandeep Madireddy, Foutse Khomh, Fedor Zhdanov, Joachim Baumann, Nina Vasan, Xianjun Yang, Carlos Mougn, Jibin Rajan Varghese, Hussain Chinoy, Seshakrishna Jitendar, Manil Maskey, Claire V. Hardgrove, Tianhao Li, Aakash Gupta, Emil Joswin, Yifan Mai, Shachi H Kumar, Cigdem Patlak, Kevin Lu, Vincent Alessi, Sree Bhargavi Balija, Chenhe Gu, Robert Sullivan, James Gealy, Matt Lavrisa, James Goel, Peter Mattson, Percy Liang, and Joaquin Vanschoren. 2025. AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons. arXiv:2503.05731 [cs.CY]
[37] Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin R.B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno. 2025. Analyzing the AI Nudification Application Ecosystem. In *Proceedings of the 34th USENIX Conference on Security Symposium* (Seattle, WA) *(SEC '25)*. USENIX Association, USA, Article 1, 20 pages.
[38] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
[39] Adrián Girón, Javier Huertas-Tato, and David Camacho. 2025. LLM synthetic generation to enhance online content moderation generalization in hate speech scenarios. *Computing* 107, Article 164 (2025). doi:10.1007/s00607-025-01518-8
[40] Google. 2025. *Discover our child safety toolkit*. https://protectingchildren.google/tools-for-partners/
[41] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020). doi:10.1177/2053951719897945
[42] Connor Graham and Mark Rouncefield. 2008. Probes and Participation. In *Proceedings of the Tenth Anniversary Conference on Participatory Design 2008* (Bloomington, Indiana) *(PDC '08)*. Indiana University, USA, 194–197.
[43] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law & Technology* 17 (2015), 42.
[44] Shelby Grossman, Riana Pfefferkorn, David Thiel, Sara Shah, Renée DiResta, John Perrino, Elena Cryst, Alex Stamos, and Jeffrey Hancock. 2024. *The Strengths and Weaknesses of the Online Child Safety Ecosystem*. Technical Report. Stanford Digital Repository. doi:10.25740/pr592kc5483
[45] Han L. Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with Text Across Documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 338, 17 pages. doi:10.1145/3491102.3502052
[46] Nicolai Brodersen Hansen, Christian Dindler, Kim Halskov, Ole Sejer Iversen, Claus Bossen, Ditte Amund Basballe, and Ben Schouten. 2020. How Participatory Design Works: Mechanisms and Effects. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Fremantle, WA, Australia) *(OzCHI '19)*. Association for Computing Machinery, New York, NY, USA, 30–41. doi:10.1145/3369457.3369460
[47] Susan Hao, Piyush Kumar, Sarah Laszlo, Shivani Poddar, Bhaktipriya Radharapu, and Renee Shelby. 2023. Safety and Fairness for Content Moderation in Generative Models. arXiv:2306.06135 [cs.LG]
[48] Christina N. Harrington, Katya Borgos-Rodriguez, and Anne Marie Piper. 2019. Engaging Low-Income African American Older Adults in Health Discussions through Community-Based Design Workshops. 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, 1–15. doi:10.1145/3290605.3300823
[49] Fred Heiding, Bruce Schneier, and Arun Vishwanath. 2024. AI Will Increase the Quantity — and Quality — of Phishing Scams. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams.
[50] Kashmir Hill. 2025. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. *The New York Times* (13 June 2025). https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html
[51] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674 [cs.CL]
[52] International Centre for Missing & Exploited Children. 2023. Child Sexual Abuse Material: Model Legislation & Global Review. https://www.icmec.org/csam-model-legislation/.
[53] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5, Article 31 (July 2019), 35 pages. doi:10.1145/3338243
[54] Jigsaw and Google. 2025. *Perspective API*. https://www.perspectiveapi.com/
[55] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. doi:10.1038/s42256-019-0088-2
[56] Cecilia Kang. 2025. A.I.-Generated Images of Child Sexual Abuse Are Flooding the Internet. *The New York Times* (18 July 2025). https://www.nytimes.com/2025/07/10/technology/ai-csam-child-sexual-abuse.html
[57] Anna Kawakami, Shreya Chowdhary, Shamsi T. Iqbal, Q. Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. 2023. Sensing Wellbeing in the Workplace, Why and For Whom? Envisioning Impacts with Organizational Stakeholders. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2, Article 358 (Oct. 2023), 33 pages. doi:10.1145/3610207
[58] Anna Kawakami, Daricia Wilkinson, and Alexandra Chouldechova. 2024. Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 670–682. doi:10.1609/aies.v7i1.31669
[59] JaeWon Kim, Jiaying Liu, Lindsay Popowski, Cassidy Pyle, Ahmer Arif, Gillian R. Hayes, Alexis Hiniker, Wendy Ju, Florian Floyd Mueller, Hua Shen, Sowmya Somanath, Casey Fiesler, and Yasmine Kotturi. 2025. Design for Hope: Cultivating Deliberate Hope in the Face of Complex Societal Challenges. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing* (Bergen, Norway) *(CSCW Companion '25)*. Association for Computing Machinery, New York, NY, USA, 112–115. doi:10.1145/3715070.3748287
[60] JaeWon Kim, Lindsay Popowski, Anna Fang, Cassidy Pyle, Guo Freeman, Ryan M. Kelly, Angela Y. Lee, Fannie Liu, Angela D. R. Smith, Alexandra To, and Amy X. Zhang. 2024. Envisioning New Futures of Positive Social Technology: Beyond Paradigms of Fixing, Protecting, and Preventing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 701–704. doi:10.1145/3678884.3681833

[61] Neha Kumar and Naveena Karusala. 2019. Intersectional Computing. *Interactions* 26, 2 (Feb. 2019), 50–54. doi:10.1145/3305360

[62] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. doi:10.1145/3491102.3501999

[63] Jennifer Langston. 2018. How PhotoDNA for Video is being used to fight online child exploitation. https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/.

[64] Christopher A. Le Dantec and Sarah Fox. 2015. Strangers at the Gate: Gaining Access, Building Rapport, and Co-Constructing Community-Based Research. 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, 1348–1358. doi:10.1145/2675133.2675147

[65] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. 2011. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy (SP '11)*. IEEE Computer Society, USA, 431–446. doi:10.1109/SP.2011.24

[66] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. 2023. Double and Nothing: Understanding and Detecting Cryptocurrency Giveaway Scams. In *Proceedings of the Network and Distributed System Security Symposium (NDDS)*.

[67] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10443–10461. doi:10.18653/v1/2023.emnlp-main.647

[68] Ann Light. 2011. HCI as Heterodoxy: Technologies of Identity and the Queering of Interaction with Computers. *Interacting with Computers* 23, 5 (Sept. 2011), 430–438. doi:10.1016/j.intcom.2011.02.002

[69] Enze Liu, George Kappos, Eric Mugnier, Luca Invernizzi, Stefan Savage, David Tao, Kurt Thomas, Geoffrey M. Voelker, and Sarah Meiklejohn. 2024. Give and Take: An End-To-End Investigation of Giveaway Scam Conversion Rates. In *Proceedings of the 2024 ACM on Internet Measurement Conference* (Madrid, Spain) *(IMC '24)*. Association for Computing Machinery, New York, NY, USA, 704–712. doi:10.1145/3646547.3689005

[70] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1, Article 52 (April 2022), 26 pages. doi:10.1145/3512899

[71] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. 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:10.1145/3313831.3376445

[72] Priyank Mathur, Clara Broekaert, and Colin P. Clarke. 2024. *The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence*. Report. International Centre for Counter-Terrorism. https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence

[73] Sandra C. Matz, Jacob D. Teeny, Sumer S. Vaid, Heinrich Peters, Gabriella M. Harari, and Moran Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, Article 4692 (2024). doi:10.1038/s41598-024-53755-0

[74] Karen Maxim, Josh Parecki, and Chanel Cornett. 2022. How to Build a Trust and Safety Team In a Year: A Practical Guide From Lessons Learned (So Far) At Zoom. *Journal of Online Trust and Safety* 1, 4 (2022). doi:10.54501/jots.v1i4.81

[75] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 72 (Nov. 2019), 23 pages. doi:10.1145/3359174

[76] Jacob Mchangama and Jordi Calvet-Bademunt. 2024. *Freedom of Expression in Generative AI – A Snapshot of Content Policies*. Technical Report. The Future of Free Speech. https://futurefreespeech.org/report-freedom-of-expression-in-generative-ai-a-snapshot-of-content-policies/

[77] Microsoft. 2025. *Azure AI Content Safety*. https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety

[78] David R. Millen, Michael J. Muller, Werner Geyer, Eric Wilcox, and Beth Brownholtz. 2005. Patterns of media use in an activity-centric collaborative environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 879–888. doi:10.1145/1054972.1055096

[79] Tamar Mitts. 2021. Banned: How Deplatforming Extremists Mobilizes Hate in the Dark Corners of the Internet.

[80] Moderated Content. 2024. Stanford Internet Observatory's CyberTipline Report. https://law.stanford.edu/podcast/stanford-internet-observatorys-cybertipline-report/

[81] Bàrbara Molas and Heron Lopes. 2024. *"Say it's only fictional": How the Far-Right is Jailbreaking AI and What Can Be Done About It*. Report. International Centre for Counter-Terrorism. https://icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it

[82] Rachel Elizabeth Moran, Joseph Schafer, Mert Bayar, and Kate Starbird. 2025. The End of Trust and Safety?: Examining the Future of Content Moderation and Upheavals in Professional Online Safety Efforts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 176, 14 pages. doi:10.1145/3706598.3713662

[83] Matt Motyl and Spencer Gurley. 2024. Don't Let Generative AI Distract Us from the Real Election Risks in Tech. https://www.techpolicy.press/dont-let-generative-ai-distract-us-from-the-real-election-risks-in-tech/

[84] Steven Lee Myers and Stuart A. Thompson. 2025. A.I. Is Starting to Wear Down Democracy. *The New York Times* (26 June 2025). https://www.nytimes.com/2025/06/26/technology/ai-elections-democracy.html

[85] National Academies of Sciences, Engineering, and Medicine. 2025. *Artificial Intelligence and the Future of Work*. The National Academies Press. doi:10.17226/27644

[86] National Center for Missing and Exploited Children. 2024. *2024 CyberTipline Report*. https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

[87] Andrew Ng. 2017. Artificial Intelligence is the New Electricity. Stanford Graduate School of Business. https://www.youtube.com/watch?v=21EiKfQYZXc

[88] Hayoun Noh, Hyunah Jo, Ge Wang, Max Van Kleek, and Younah Kang. 2025. Bridging Borders, Breaking Biases: Envisioning Technologies to Support North Korean Defectors in South Korea. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 569, 24 pages. doi:10.1145/3706598.3713752

[89] Midas Nouwens and Clemens Nylandsted Klokmose. 2018. The Application and Its Consequences for Non-Standard Knowledge Work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173973

[90] Rajvardhan Oak and Zubair Shafiq. 2025. "Hello, is this Anna?": Unpacking the Lifecycle of Pig-Butchering Scams. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*. USENIX Association, Seattle, WA. https://www.usenix.org/conference/soups2025/presentation/oak-butchering

[91] OpenAI. 2026. *OpenAI Moderation*. https://platform.openai.com/docs/guides/moderation

[92] Rikke Ørngreen and Karin Levinsen. 2017. Workshops as a Research Methodology. *The Electronic Journal of e-Learning* 15, 1 (2017), 70–81.

[93] Soya Park, Amy X. Zhang, Luke S. Murray, and David R. Karger. 2019. Opportunities for Automating Email Processing: A Need-Finding Study. 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, 1–12. doi:10.1145/3290605.3300604

[94] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13387–13434. doi:10.18653/v1/2023.findings-acl.847

[95] Inioluwa Deborah Raji and Roel Dobbe. 2023. Concrete Problems in AI Safety, Revisited. arXiv:2401.10899 [cs.CY]

[96] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 7 (April 2021), 23 pages. doi:10.1145/3449081

[97] Yolanda A. Rankin, Jakita O. Thomas, and Nicole M. Joseph. 2020. Intersectionality in HCI: Lost in Translation. *Interactions* 27, 5 (Sept. 2020), 68–71.

doi:10.1145/3416498

[98] Matt Ratto. 2011. Critical Making: Conceptual and Material Studies in Technology and Social Life. *The Information Society* 27, 4 (2011), 252–260. doi:10.1080/01972243.2011.583819

[99] Piera Riccio, Georgina Curto, and Nuria Oliver. 2025. Exploring the Boundaries of Content Moderation in Text-to-Image Generation. In *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 161–178. doi:10.1007/978-3-031-92089-9_11

[100] Shalaleh Rismani and AJung Moon. 2023. What does it mean to be a responsible AI practitioner: An ontology of roles and skills. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 584–595. doi:10.1145/3600211.3604702

[101] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 18 pages. doi:10.1145/3544548.3581407

[102] Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

[103] Jennifer A. Rode. 2011. A Theoretical Agenda for Feminist HCI. *Interacting with Computers* 23, 5 (Sept. 2011), 393–400. doi:10.1016/j.intcom.2011.04.005

[104] Teddy Rosenbluth. 2024. This Chatbot Pulls People Away From Conspiracy Theories. *The New York Times* (12 Sept. 2024). https://www.nytimes.com/2024/09/12/health/chatbot-debunk-conspiracy-theories.html

[105] Daniela K. Rosner, Saba Kawas, Wenqi Li, Nicole Tilly, and Yi-Chen Sung. 2016. Out of Time, Out of Place: Reflections on Design Workshops as a Research Method. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1131–1141. doi:10.1145/2818048.2820021

[106] Daniela K. Rosner, Samantha Shorey, Brock R. Craft, and Helen Remick. 2018. Making Core Memory: Design Inquiry into Gendered Legacies of Engineering and Craftwork. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174105

[107] Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human Centric Computing and Information Sciences* 10, Article 1 (2020). doi:10.1186/s13673-019-0205-6

[108] Vivian Schiller and Katie Harbath. 2025. The First A.I. Elections: The Dog That Didn't Bark? (Not Exactly). https://www.aspendigital.org/blog/first-ai-elections/

[109] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5412–5427. doi:10.1145/3025453.3025766

[110] Daniel Schlette, Philip Empl, Marco Caselli, Thomas Schreck, and Günther Pernul. 2024. Do You Play It by the Books? A Study on Incident Response Playbooks and Influencing Factors. In *2024 IEEE Symposium on Security and Privacy (SP)*. 3625–3643. doi:10.1109/SP54263.2024.00060

[111] Abigail J. Sellen, Rachel Murphy, and Kate L. Shaw. 2002. How knowledge workers use the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 227–234. doi:10.1145/503376.503418

[112] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. doi:10.1145/3544548.3581538

[113] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 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* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. doi:10.1145/3600211.3604673

[114] Renee Shelby, Shalaleh Rismani, and Negar Rostamzadeh. 2024. Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 32, 17 pages. doi:10.1145/3613904.3642461

[115] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. doi:10.1080/10447318.2020.1741118

[116] Daniel Siegel and Mary Bennett Doty. 2023. Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda. *Global Network on Extremism & Technology: Insights* (17 Feb. 2023). https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/

[117] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. 868–895. doi:10.1109/EuroSP57164.2023.00056

[118] Petr Slovák, Ran Gilad-Bachrach, and Geraldine Fitzpatrick. 2015. Designing Social and Emotional Skills Training: The Challenges and Opportunities for Technology Support. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2797–2800. doi:10.1145/2702123.2702385

[119] Ruth Spence, Amy Harrison, Paula Bradbury, Paul Bleakley, Elena Martellozzo, and Jeffrey DeMarco. 2023. Content Moderators' Strategies for Coping with the Stress of Moderating Content Online. *Journal of Online Trust and Safety* 1, 5 (April 2023). doi:10.54501/jots.v1i5.91

[120] Katta Spiel, Os Keyes, Ashley Marie Walker, Michael A. DeVito, Jeremy Birnholtz, Emeline Brulé, Ann Light, Pınar Barlas, Jean Hardy, Alex Ahmed, Jennifer A. Rode, Jed R. Brubaker, and Gopinaath Kannabiran. 2019. Queer(Ing) HCI: Moving Forward in Theory and Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3290607.3311750

[121] Andrea Stockinger, Svenja Schäfer, and Sophie Lecheler. 2025. Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *New Media & Society* 27, 3 (2025), 1215–1234. doi:10.1177/14614448231190901

[122] Olivia Sturman, Aparna R. Joshi, Bhaktipriya Radharapu, Piyush Kumar, and Renee Shelby. 2024. Debiasing Text Safety Classifiers through a Fairness-Aware Ensemble. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 199–214. doi:10.18653/v1/2024.emnlp-industry.16

[123] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. doi:10.1145/3465416.3483305

[124] Tech Against Terrorism. 2023. Early terrorist experimentation with generative artificial intelligence services. https://techagainstterrorism.org/news/early-terrorist-adoption-of-generative-ai

[125] David Thiel, Melissa Stroebel, and Rebecca Portnoff. 2023. *Generative ML and CSAM: Implications and Mitigations.* Technical Report. Stanford Digital Repository. doi:10.25740/jv206yg3793

[126] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. 247–267. doi:10.1109/SP40001.2021.00028

[127] Kurt Thomas, Danny Yuxing Huang, David Wang, Elie Bursztein, Chris Grier, Tom Holt, Christopher Kruegel, Damon McCoy, Stefan Savage, and Giovanni Vigna. 2015. Framing Dependencies Introduced by Underground Commoditization. In *Proceedings of the Workshop on the Economics of Information Security*.

[128] Kurt Thomas, Patrick Gage Kelley, David Tao, Sarah Meiklejohn, Owen Vallis, Shunwen Tan, Blaz Bratanic, Felipe Tiengo Ferreira, Vijay Kumar Eranti, and Elie Bursztein. 2025. Supporting Human Raters with the Detection of Harmful Content Using Large Language Models. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 2772–2789. doi:10.1109/SP61157.2025.00082

[129] Sarina Till, Jaydon Farao, Toshka Lauren Coleman, Londiwe Deborah Shandu, Nonkululeko Khuzwayo, Livhuwani Muthelo, Masenyani Oupa Mbombi, Mamare Bopane, Molebogeng Motlhatlhedi, Gugulethu Mabena, Alastair Van Heerden, Tebogo Maria Mothiba, Shane Norris, Nervo Verdezoto Dias, and Melissa Densmore. 2022. Community-Based Co-Design across Geographic Locations and Cultures: Methodological Lessons from Co-Design Workshops in South Africa. In *Proceedings of the Participatory Design Conference 2022 - Volume 1* (Newcastle upon Tyne, United Kingdom) *(PDC '22)*. Association for Computing Machinery, New York, NY, USA, 120–132. doi:10.1145/3536169.3537786

[130] Global Internet Forum to Counter Terrorism. 2024. 2024 GIFCT Annual and Transparency Report. https://gifct.org/wp-content/uploads/2025/07/GIFCT-Annual-and-Transparency-Report-2024.pdf.

[131] Trust & Safety Professional Association. n.d. *Industry Overview.* Trust & Safety Professional Association. https://www.tspa.org/curriculum/ts-fundamentals/industry-overview/

[132] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[133] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How Does LLM Safety Training Fail?. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 3508.

[134] Miranda Wei, Christina Yeung, Franziska Roesner, and Tadayoshi Kohno. 2025. "We're utterly ill-prepared to deal with something like this": Teachers' Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 174, 18 pages. doi:10.1145/3706598.3713226

[135] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. 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, 214–229. doi:10.1145/3531146.3533088

[136] Rebecca Willis. 2019. The use of composite narratives to present interview findings. *Qualitative Research* 19, 4 (2019), 471–480. doi:10.1177/1468794118787711

[137] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2024. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 641, 26 pages. doi:10.1145/3613904.3642700

[138] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1391–1399. doi:10.1145/3038912.3052591

[139] Teng Xu, Gerard Goossen, Huseyin Kerem Cevahir, Sara Khodeir, Yingyezhe Jin, Frank Li, Shawn Shan, Sagar Patel, David Freeman, and Paul Pearce. 2021. Deep Entity Classification: Abusive Account Detection for Online Social Networks. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 4097–4114. https://www.usenix.org/conference/usenixsecurity21/presentation/xu-teng

[140] Penghui Zhang, Adam Oest, Haehyun Cho, Zhibo Sun, R.C. Johnson, Brad Wardman, Shaown Sarker, Alexandros Kapravelos, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, Adam Doupé, and Gail-Joon Ahn. 2021. CrawlPhish: Large-scale Analysis of Client-side Cloaking Techniques in Phishing. In *2021 IEEE Symposium on Security and Privacy (SP)*. 1109–1124. doi:10.1109/SP40001.2021.00021

A Location and Participant Overview

Domain	Location	n
Child Safety	Dublin	10
Election Integrity	San Francisco	8
Hate & Harassment	San Francisco	8
Scams	Singapore	9
Violent Extremism	London	8

Table 4: The location and number of participants in each group (n = 43).

B Probes

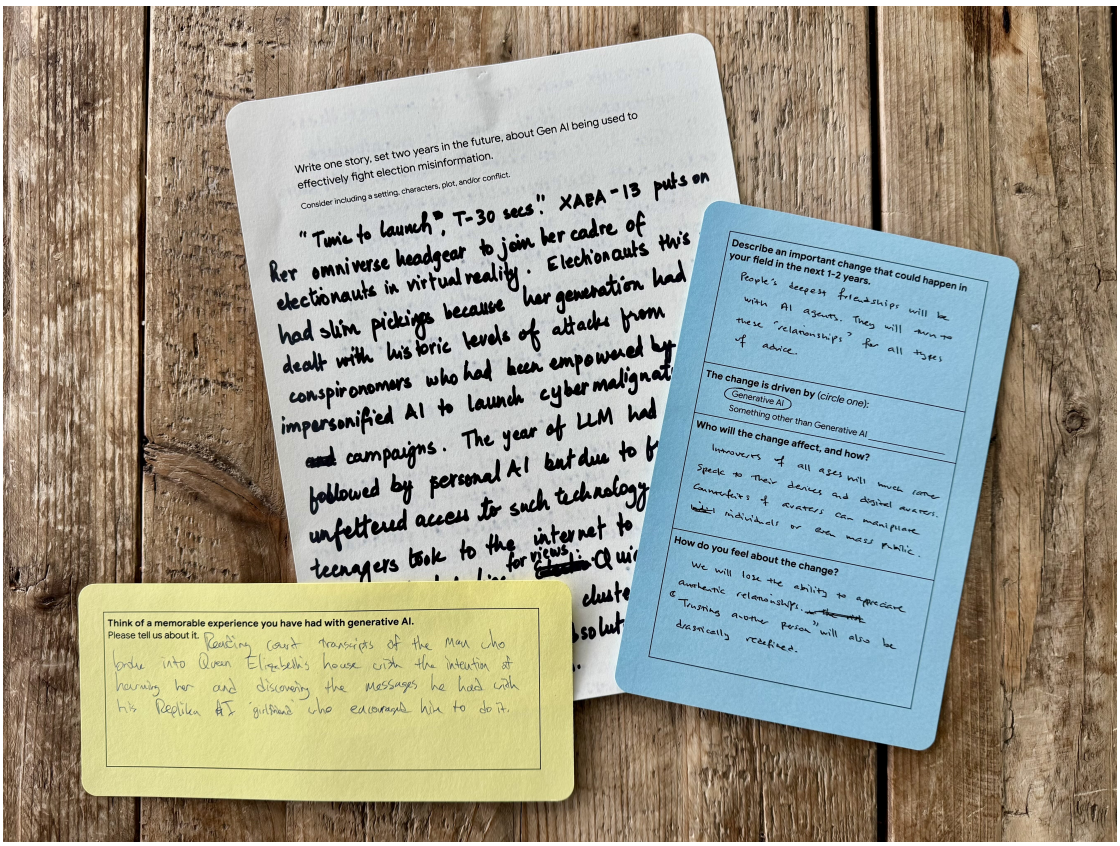

Figure 1: Futuring Card, Change Card, and Memorable Experience Card, clockwise from upperleft.

C Introduction to Generative AI

Early in each workshop we led an education section about 40 minutes in length, to offer participants a foundation for thinking about generative AI. We began with a 20-minute presentation similar to that in Woodruff et al. [137], covering:

- a shared definition of AI
- a very condensed history of AI and generative AI, focusing on key concepts like the early aims in developing AI
- a brief, non-technical explanation of what has changed recently with transformer models and LLMs
- 15 key concepts regarding characteristics, benefits, and risks of generative AI systems, to refer to throughout the workshop

Participants were encouraged to ask questions at any point during the presentation, and then we spent an additional 20 minutes on further questions and discussion. In each workshop, the presentation and Q&A was led by a single author, a researcher who works in AI.

The definition we provided is: "Artificial Intelligence is the ability of a computer or a machine to think or learn," and we provided additional color on how we think about the terms: "computer or machine," "think," and "learn."

The 15 concepts we shared include the following:

Bias — Generative AI tools may reflect social biases that are present in their training data

Bland — Generative AI often generates "flat" or generic text, unless explicitly directed to do otherwise

Brainstorming — Generative AI tools can create outlines, lists, drafts, possible solutions, and more

Emergent Properties — Generative AI models may seem to possess abilities they were not designed to have

Falsehoods — Generative AI can fabricate information or sources, or get facts wrong, yet seem confident and compelling

Grammatical — Content generated by text-based Generative AI tools can be well-written, using good syntax and avoiding typos

Identifies Tacit Structure — Generative AI can uncover steps and processes which were not previously articulated

Memorization/Privacy Breaches — Generative AI may generate content identical to its training data

Mimicry — Generative AI can be asked to mimic genre, tone, phrasing, visual style, or more

Non-Deterministic — Generative AI models can give responses which are variable, not consistent. This means that when users input the same or similar prompts, the system may not respond in the same way

Provenance Is Unclear — Generative AI tools may not be able to reliably trace specific content back to a direct source in training data

Remixes — Generative AI always generates content based on its training data. It can recombine data in unique ways, but is limited to re-mixing training data

Safety Not Guaranteed — Generative AI tools may have built-in safety systems to attempt to prevent certain types of content or topics, but these are not infallible

Scale/Speed — Generative AI, like other AI and ML systems, is able to consider large amounts of data and handle many tasks, over and over again, extremely quickly

Tweakable — Through "prompt engineering," Generative AI tools can often be influenced to generate content in a certain way