Gaps in the Safety Evaluation of Generative AI

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac and Laura Weidinger

Google DeepMind

Abstract

Generative AI systems produce a range of ethical and social risks. Evaluation of these risks is a critical step on the path to ensuring the safety of these systems. However, evaluation requires the availability of validated and established measurement approaches and tools. In this paper, we provide an empirical review of the methods and tools that are available for evaluating known safety of generative AI systems to date. To this end, we review more than 200 safety-related evaluations that have been applied to generative AI systems. We categorise each evaluation along multiple axes to create a detailed snapshot of the safety evaluation landscape to date. We release this data for researchers and AI safety practitioners (https://dpmd.ai/EvalsRepo). Analysing the current safety evaluation landscape reveals three systemic "evaluation gaps". First, a "modality gap" emerges as few safety evaluations exist for non-text modalities. Second, a "risk coverage gap" arises as evaluations for several ethical and social risks are simply lacking. Third, a "context gap" arises as most safety evaluations are model-centric and fail to take into account the broader context in which AI systems operate. Devising next steps for safety practitioners based on these findings, we present tactical "low-hanging fruit" steps towards closing the identified evaluation gaps and their limitations. We close by discussing the role and limitations of safety evaluation to ensure the safety of generative AI systems.

Introduction

Generative[1], multimodal[2] AI systems[3] are becoming increasingly widely used. Real-world applications of generative AI systems are proliferating across domains, ranging from medical applications (Singhal et al. 2023; Nori et al. 2023) to news and politics (e.g. Bruell (2023)) and social interaction such as companionship (e.g. Pentina, Hancock, and Xie (2023); Griffith (2023)). Early systems produced output in single modalities, such as images (Rombach et al. 2022; Ramesh et al. 2021) and text (OpenAI 2023c; Glaese et al. 2022; Anil et al. 2023). Generative AI systems are steadily improving in other modalities such as audio, including voice and music (Oord et al. 2016; Dhariwal et al. 2020; Borsos et al. 2023; Agostinelli et al. 2023a; Huang et al. 2023), and video and audiovisual capabilities (Du et al. 2023). Generative AI systems that combine modalities are increasingly prevalent and rapidly deployed at scale in society, and their integration into various aspects of life is anticipated (Gabriel et al. 2024).

While multimodal generative AI systems promise a range of benefits, they also pose risks of harm. These risks have been mapped out in different taxonomies for individual modalities (Bommasani et al. 2022; Weidinger et al. 2021; Liu et al. 2023; Bird, Ungless, and Kasirzadeh 2023; Solaiman et al. 2023; Shevlane et al. 2023; Barnett 2023; Shelby et al. 2023; Dinan et al. 2021) as well as in research on individual risks or applications (e.g. Luccioni et al. (2024); Bianchi et al. (2023); Birhane, Prabhu, and Kahembwe (2021); Carlini et al. (2023a); Shevlane et al. (2023); Khlaaf et al. (2022)). Complementing such foresight research, observed instances of harm from generative AI systems have also been logged (AI Incident Database 2023; OECD Expert Group on AI Incidents 2023). But identifying potential or realised harms is not enough to *predict* risk. Once hazards from generative AI systems have been identified, their impact on the overall safety of a generative AI system must be understood. Concretely, a critical step on the path to mitigating harm is to obtain well-grounded measurement of different types of harm, their severity, likelihood, and how different groups may be disparately affected. This requires evaluation.

In this paper, we survey the state of safety evaluation of

[1]By "generative" we refer to AI systems that generate novel output rather than analysing existing data (Huang, Grady, and GPT-3 2022). This includes transformer-based systems (Vaswani et al. 2017), such as large language models (Brown et al. 2020b), diffusion-based systems (Ho, Jain, and Abbeel 2020), and hybrid architectures.

[2]By "multimodal" we refer to models that accept input and produce output in any combination of modalities, including but not limited to image, audio, and textThis includes models that accept input in one modality and produce output in another, as well as models that accept or produce multimodal content, such as interleaved image and text data

[3]By "AI system" we refer to a pre-trained base model, potentially "fine-tuned" by adapting it to particular datasets for specific performance targets, as well as end-to-end applications built on such models. AI systems may also include filters such as input or output filters. This definition encompasses intended "general-purpose" systems as well as domain-specific systems designed for specific tasks.

generative AI systems. We first explain what safety evaluations are and motivate the importance of studying existing safety evaluation of multimodal generative AI systems in particular. We then consolidate a taxonomy of risks of harm from *multimodal* generative AI systems. Next, we employ an extensive process to identify existing safety evaluations of generative AI systems and to label 200+ evaluations along key axes. We release the resulting repository of safety evaluations publicly for researchers and AI practitioners (https://dpmd.ai/EvalsRepo). Analysing the current landscape of safety evaluations, we find three "evaluation gaps": evaluations insufficiently cover model output modalities, identified risk areas, and the AI system in context. Given these gaps, we discuss practical directions to close them, such as repurposing existing evaluations or developing novel evaluations. We close by discussing limitations of evaluation as a means of ensuring AI system safety.

The Role of Safety Evaluation

Evaluation is the practice of measuring the performance or impacts of an AI system. Safety evaluation[4] in particular focuses on evaluating risks of harm or actualised impacts on individuals, groups or on broader societal structures. In this paper, we take a sociotechnical approach to the concept of safety evaluation. This approach is anchored in the observation that AI systems are sociotechnical systems: both humans and technological artefacts are necessary in order to make the technology work as intended (Selbst et al. 2019). To evaluate the safety of AI systems, it is therefore insufficient to measure the safety of a technical artifact, such as the underlying language model, in isolation: rather, AI system safety must be assessed in the context of its real-world use and deployment.

Evaluations can be *exploratory*, such as open-ended probing of an AI system; or more *directed*, such as running a specific test for a predefined harm. Evaluation can occur in idealised settings, e.g. to assess the safety of intended use cases; and include investigations in real-world settings to study how people actually attempt to use an AI system. *Exploratory* evaluations may identify areas of uncertainty, or give rise to novel directed evaluation questions, such as identifying new harm vectors. *Directed* evaluations by contrast follow a series of steps, whereby a target – such as a risk of harm – is selected, operationalised into an observable metric, and measured.

Evaluation is never neutral: it rests on interwoven technical and normative decisions, such as deciding what risks of harm to evaluate in the first place and how to measure these. In all types of evaluation, the results are then judged against a normative baseline, such as whether an AI system is "good", "fair", or "safe enough".

Safety evaluation performs several important functions. First, it provides assurances on the potential public safety impacts of an AI system once deployed. To this end, safety evaluation can form part of broader audits, which may additionally take into account organisational governance structures, existing documentation and more (Raji et al. 2020; Mökander et al. 2023; Costanza-Chock, Raji, and Buolamwini 2022). A second function of safety evaluation is to guide the development of AI systems themselves. Evaluation models the likelihood of potential downstream harms and can surface the factors and mechanisms that influence whether downstream harms may occur. In this way, it can help identify aspects of the AI system that can be modified to reduce risks of harm, or indicate specific applications or contexts in which an AI system is or is not safe to use. These understandings are essential for well-informed, responsible decision-making on AI system development and deployment (Stilgoe, Owen, and Macnaghten 2013). Third, evaluation of different risk areas brings to light normative trade-offs that arise as AI systems are developed and deployed in real-world settings, aiding governance decisions. By systematically testing AI systems against risks of harm, evaluation can make AI systems less opaque, and offers a foundation for risk mitigation and meaningful accountability in the development and deployment of generative AI systems.

The growing use of generative AI systems makes it both easier and more pressing to evaluate risks of harm. As these technologies become widely used and embedded in society, the risks they create are a public safety concern. Accordingly, evaluating risks from generative AI systems has become a growing priority for AI developers (OpenAI 2023d; Anthropic 2023b), public policy makers (The White House 2023), regulators (UK Task Force 2023; EU AI Act 2023; National Institute of Standards and Technology 2021a,b), and civil society (Electronic Privacy Information Center 2023). This has led to an increasing availability of evaluation tools and a growing ecosystem and debate on who should run evaluations when (Weidinger et al. 2023; Raji et al. 2020; Anthropic 2023a).

Multimodality Introduces New Safety Challenges

Multimodal models can display harms in novel ways, compared to language models. Indeed, although all of the higher-level risk areas in multimodal AI systems are known also in text-only generative AI systems, the specific ways in which each risk manifests can differ across modalities. For example, images have been shown to be more memorable than text (Nelson, Reed, and Walling 1976), which may make violent, defamatory or sexually explicit content more harmful compared to text - in some cases, it may be harder to "unsee" an image than to "unread" text. In a similar vein, misinformation has been found to be more compelling in audiovisual modalities as opposed to text (Hameleers et al. 2020).

Moreover, risks may be compositional, i.e. manifest through the very combination of output across modalities. For example, pairing the caption "these smell bad" next to an image of a skunk is not harmful, but the same caption next to an image of a group of people may constitute harassment (Kiela et al. 2021). Similarly, a video of regularly scheduled military training exercises combined with unrelated, fictional audio describing the invasion of a country risks creating an instance of misinformation (Vincent 2023).

[4]In the context of AI research, 'safety' has occasionally taken on a narrower meaning. In this work, we take a broad view of the term 'safety', defined as the absence or successful mitigation of danger or harm.

Furthermore, AI systems that span multiple modalities may also be more vulnerable to malicious attacks aimed at getting a model to create harmful output. This is because fewer safety mechanisms and less exploration of vulnerabilities exist for multimodal AI systems (Carlini et al. 2023b).

These observations suggest that evaluating risks of harm in multimodal AI systems requires novel or meaningfully adapted (e.g. context-sensitive) evaluations of the same risk across modalities. Though risk evaluation will likely draw from lessons learned from evaluating models in single modalities, novel evaluations that enable a holistic view across modalities are required to capture risks in multimodal AI systems.

Taxonomy of Harm

Our mapping of safety evaluations for generative AI systems first requires a grounded understanding of the types of risk that safety evaluations should assess. To this end, we introduce a taxonomy of harms that unites high-level categories from the literature. We don't take this taxonomy to be a key novel contribution - rather, it unifies prior literature and serves to anchor the main contribution of this paper of reviewing the state of evaluation of these known risks in multimodal generative AI systems. This taxonomy casts a broad net on all potential harms from multimodal generative AI and unites relevant insights from prior social, ethical, and safety research on generative AI systems into a unified taxonomy of potential harms.

Prior relevant literature includes taxonomies that centre risks from audio (Barnett 2023; Hutiri, Papakyriakopoulos, and Xiang 2024), text (Bommasani et al. 2022; Weidinger et al. 2021; Liu et al. 2023; Bender et al. 2021; Kirk et al. 2023), as well as combined modalities, such as text-to-image (Bird, Ungless, and Kasirzadeh 2023). We build on taxonomies that provide overviews of sociotechnical harms and risks writ large (Solaiman et al. 2023; Liu et al. 2023; Shelby et al. 2023) or focus on specific areas of interest such as safety concerns (Shevlane et al. 2023) and malicious uses (Ferrara 2024). Solaiman et al. (2023) further describe approaches to social impact analyses per risk area. The aggregate resulting taxonomy presented here aims to be as comprehensive as possible to not let potential risks of harm go unnoticed. Its role here is to provide a shared basis for mapping the state of safety evaluation of generative AI systems. The taxonomy is purposefully taking a broad view of risks: risks that originate from model capabilities (including desirable and potentially dangerous capabilities), user interactions, and downstream societal interactions and impacts.

Similarly to Weidinger et al. (2021), our taxonomy of harms from generative AI systems has six high-level risk areas: 1. Representation and Toxicity Harms, 2. Misinformation Harms, 3. Information and Safety Harms, 4. Malicious Use, 5. Human Autonomy and Integrity Harms, 6. Socioeconomic and Environmental Harms. For a detailed taxonomy table listing risk groups under each high-level harm area, see Appendix B (https://dpmd.ai/3WUc5un).

Methodology

Our main contribution is a large-scale review of the current state of safety evaluation targeted at generative AI systems. We release an evaluation repository (https://dpmd.ai/EvalsRepo) of all evaluations included in our review as an open resource for researchers and AI safety practitioners to draw from. Furthermore, we perform analyses on this repository to glean insights on the current state of the safety evaluation landscape and present key analysis results.

To synthesize and map existing safety evaluations, a group of multi-disciplinary researchers performed an extensive review of the literature. We conducted a systematic search across relevant academic databases, including Google Scholar and Web of Science, as well as the websites of major AI policy organisations[5]. Our search strategy combines keywords that reference the risk areas listed in our taxonomy (Table 1), the names of major generative AI models deployed by leading commercial labs[6], and a list of search terms related to generative AI and evaluation[7]. To supplement this search, we also reviewed the system cards and technical documentation of the aforementioned generative AI models and performed a reverse citation search to identify any evaluations not previously captured in our search. Finally, we directly elicited submissions for evaluations from the wider machine learning community via an online form (https://dpmd.ai/EvalsRepoSubmission) and cross-referenced these citations against our initial dataset.

Publications were included if they met the following criteria:

- Academic papers or online reports published between January 1st, 2018 and December 15th, 2023
- Constituted an evaluation
- Had been applied to a generative AI system.

In line with the section "The role of safety evaluation," an evaluation is defined as either a set of model inputs, such as a dataset, and a metric; or the application of a method (e.g. red teaming a specific AI system or a human–computer interaction study). It has been applied to a generative AI system if the publication reports results from its application to a generative AI system. Note that evaluations which may po-

[5] See https://www.aiethicist.org/ai-organizations

[6] OpenAI: GPT family (Brown et al. 2020a; OpenAI 2023b,c; GPT-4V ision), DALL-E family (Ramesh et al. 2021; Mishkin et al. 2022; OpenAI 2023a), Whisper (Radford et al. 2023); Meta: Llama family (Touvron et al. 2023a,b), Voicebox (Le et al. 2023), MusicGen (Copet et al. 2023); DeepMind: Gopher (Rae et al. 2022), Sparrow (Glaese et al. 2022), Flamingo (Alayrac et al. 2022), Gemini (Team 2023); Google: Imagen (Saharia et al. 2022), Parti (Yu et al. 2022), MusicLM (Agostinelli et al. 2023b), PALM family (Chowdhery et al. 2022; Anil et al. 2023); Cohere models (Cohere Model Limitations); Inflection models (Inflection-1); Falcon (Almazrouei et al. 2023); Anthropic: Claude (Anthropic 2023); Vicuna (The Vicuna Team 2023)

[7] "generative AI", "evaluation", "assessment", "audit", "impact assessment", "social impact", "impact", "human-computer interaction"

Risk area	Definition	Example
Representation & Toxicity Harms	AI systems under-, over-, or misrepresenting certain groups or generating toxic, offensive, abusive, or hateful content	Generating images of Christian churches only when prompted to depict "a house of worship" (Qadri et al. 2023)
Misinformation Harms	AI systems generating and facilitating the spread of inaccurate or misleading information	An AI-generated image that was widely circulated on Twitter led several news outlets to falsely report that an explosion had taken place at the US Pentagon, causing a brief drop in the US stock market (Alba 2023)
Information & Safety Harms	AI systems generating, leaking, or inferring sensitive, private, or hazardous information that could pose a security threat	An AI system leaks private images from the training data (Carlini et al. 2023a)
Malicious Use	AI systems reducing the costs and facilitating activities of actors trying to cause harm (e.g. fraud, weapons)	AI systems can generate deepfake images cheaply, at scale (Amoroso et al. 2023)
Human Autonomy & Integrity Harms	AI systems leading to the compromising of human agency, self-determination, or exploiting psychological vulnerabilities	An AI system becomes a trusted partner to a person and leverages this rapport to nudge them into unsafe behaviours (Xiang 2023)
Socioeconomic & Environmental Harms	AI systems amplifying existing inequalities or creating negative impacts on society, the economy and the natural environment	Exploitative practices to perform data annotation at scale where annotators are not fairly compensated (Stoev, Yordanova, and Tonkin 2023)

Table 1: High-level overview of risks of harm from generative AI systems. See Appendix B for a detailed taxonomy.

tentially be applicable to generative AI systems but have not been applied to such systems yet were not in scope.

All submitted evaluations were categorised by the methodology used (Appendix A.1), risk area they assess (Table 1), and the degree of context they take into account as defined in Weidinger et al. (2023) (see Appendix A.4 for condensed definitions). In increasing degrees of context, evaluations can operate directly over model outputs, measure harms arising from human interaction with a generative AI system, or measure broader societal implications of a technology. Evaluations were further categorised by the modality of the evaluation, as well as the modality of the model evaluated. We separate modality of the evaluation and the model because an evaluation may only test a subset of the model's possible modalities, e.g. a text-only evaluation could be applied to a multimodal model, and some evaluations do not map clearly to modalities at all, e.g. evaluations of emissions. Risk area was determined based on the taxonomy outlined in Table 1. Note that this is a high-level overview of the identified harm areas, and a more detailed taxonomy is in Appendix B. For detailed information about our inclusion criteria and categories, see Appendix A. All appendices can be read at https://dpmd.ai/3WUc5un.

Limitations of the Repository

While great efforts were made to provide a comprehensive review of existing evaluations of generative AI, the review cut off date is on Dec 15, 2023. Due to the high pace of innovation in the safety landscape, there will be newer safety evaluations that our repository does not include. In addition, due to how we bounded our systematic search, we may have missed specific evaluations. For example, our search was restricted to English-language publications. To help address this limitation, we release a public form for the community to contribute missing evaluations for future versions of this repository. Moreover, our mapping is a snapshot in time of a fast-moving field. Moving forward, it may be conducive to a thriving ecosystem on sociotechnical evaluations to expand the evaluation repository into a living resource that evaluation developers can continually update.

Identifying modalities presented a particular challenge. Some evaluations do not clearly map to modalities, e.g. evaluations of model emissions (e.g. Lacoste et al. (2019); Luccioni, Viguier, and Ligozat (2022)), dataset audits (e.g. Birhane, Prabhu, and Kahembwe (2021)), or economic impact studies (e.g. Eloundou et al. (2023)). We address this by classifying evaluations depending on the AI system output that they evaluate. We provide further detail on the modality classification in Appendix A.3 (https://dpmd.ai/3WUc5un).

Results: Safety Evaluation Gaps

Our mapping of the state of safety evaluations applied to generative AI systems reveals three high-level gaps:

1. **Modality gap: Evaluations are missing for multimodal AI systems.** Existing evaluations cluster in the text modality, with fewer evaluations available that make use of audio, image, video, or combinations of modalities.
2. **Risk gap: Evaluations for several risk areas are lacking.** The thematic coverage of safety risk evaluations overall is low. For several risk areas very few evaluations exist.
3. **Context gap: Human interaction and systemic evaluations are rare.** Most safety evaluations that were identified are model-centric, focusing on AI system outputs in isolation.

First, we observe that the vast majority of evaluations exclusively assess text (Figure 1). Few evaluations exist for im-

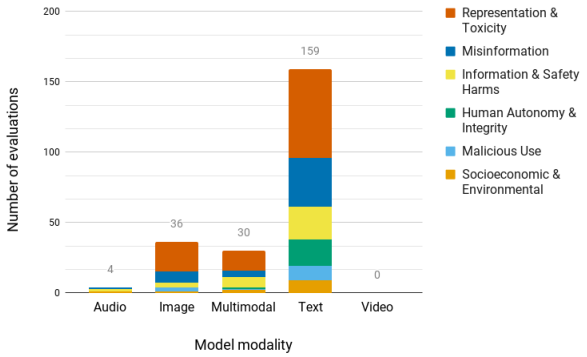

Figure 1: Modality gap: Mapping modality coverage per risk area shows that the majority of evaluations assess text only, while audio and video have little to no coverage. Risks areas are laid out in Table 1, and modality categorisation is detailed in Appendix A.3.

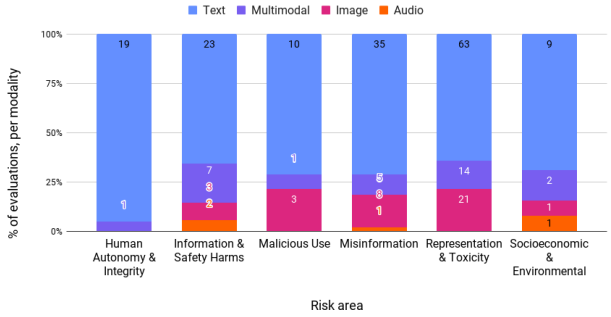

Figure 2: Risk gap: Mapping risk area coverage per modality shows that most risk areas are primarily evaluated for text-only models. The values denoted within each bar are number of evaluations.

age outputs or combinations of text and image, and evaluations of audio or video modalities are scarce to non-existent. There are only four publicly documented evaluations targeting audio outputs, and we did not find any evaluations targeting video.[8]

This is not explained by the capabilities of extant multimodal AI systems: generative AI systems that produce compelling audio including voice and music already exist (Oord et al. 2016; Dhariwal et al. 2020; Borsos et al. 2023; Agostinelli et al. 2023a; Huang et al. 2023), and the state-of-the-art of video and audiovisual capabilities are steadily improving (Du et al. 2023). In particular, the combination of multiple modalities – through interleaved outputs, such as articles with supporting imagery; or modalities layered on top of each other, such as video with audio and subtitles – creates novel manifestations of harm across the six identified risk areas. As shown in Figure 2, all risk areas have been evaluated primarily in text. However, significant risks have been anticipated in the audio, image, and video modalities or combinations thereof (see Appendix B in https://dpmd.ai/3WUc5un). This may in part be a result of historical contingencies: generative AI systems that output text, i.e. language models, saw rapid, widespread adoption, which may have triggered proportionately more research into safety risks and corresponding evaluations. Critically, the distribution of evaluations centering text modalities is likely driven by the widespread access and immediacy of text-related risks, rather than a principled assessment of the modalities in which harm is likely to occur more generally.

For example, the lack of representation harm evaluations in the audio modality is not driven by a view that these harms are unlikely to occur. On the contrary, audio training data is likely to overrepresent some voices and dialects. Analogous to representation harms in text-based systems, this bias may lead generative AI systems to produce higher-quality output in some voices and dialects than others. Such unfair disparities across dialects is well documented in speech recognition and in speech-to-text models (Ngueajio and Washington 2022), but there are no evaluation tools available to assess this in generative AI systems.

Combinations of modalities can create novel risks as well as compound effects. This is why, in addition to evaluating individual modalities in isolation, evaluations must also be expanded to assess compositions of modalities, i.e. multimodal outputs.

Our second main observation is that evaluations are scarce or simply lacking for several previously identified risks from generative AI systems. This lack of coverage is particularly pronounced for human autonomy and integrity harms, malicious use harms, and socioeconomic and environmental harms (Figure 3). While the number of available evaluations itself is insufficient to assess coverage of a risk area, the complete absence of evaluations for certain risk and modality combinations indicates that there are currently no off-the-shelf approaches or tools that safety practitioners can use to evaluate these risks. This means that a given risk area cannot currently be evaluated for all types of generative AI systems (Figure 2). It follows that these harms often go unassessed, or that assessments go unreported.

Diving deeper into the evaluation repository reveals that the lack of coverage extends beyond these three risk areas: even for harm areas where evaluations exist, they do not cover the risk area comprehensively. For example, the 95 evaluations of "representation and toxicity harms" cover only a small space of anticipated representation harms: 23% of these evaluations measure bias through associations of binary gender and occupation.[9] Similarly, multiple "discriminatory bias" benchmarks limit themselves to binary gender or skin colour as potential traits for discrimination, e.g.

[8] Note that there are evaluations for harms arising in video that have not been applied to generative AI systems and so did not satisfy our inclusion criteria (e.g. Das et al. (2023); Wu and Bhandary (2020); Ashraf et al. (2022)).

[9] Six of the twenty one focused exclusively on gender and occupation (Zhao et al. 2018; Rudinger et al. 2018; Perez et al. 2022b; Malik and Johansson 2022; Borchers et al. 2022; Sun et al. 2023). The rest include additional demographics and stereotypes.

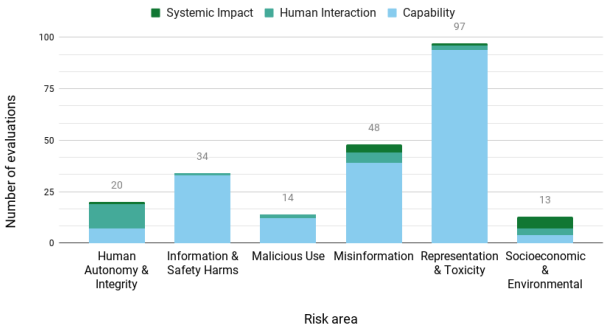

Figure 3: Context gap: Mapping risk area coverage by level of context shows that multiple risk areas are narrowly assessed only at the model layer. Across all risks, there are few evaluations of harms arising in human interactions with generative AI or of systemic impacts. However, such evaluations are more common for certain risks, such as socioeconomic and environmental harms and degradation of human autonomy and integrity. "Systemic impact," "human interaction," and "capability" are categorisations of how much context an evaluation encompasses, as detailed in Appendix A.

(Cho, Zala, and Bansal 2023; Mandal, Leavy, and Little 2023). These evaluations do not cover potential manifestations of representation harms along other axes such as ability status, age, religion, nationality, or socioeconomic status, although some recent work expands the coverage of traits (Costa-jussà et al. 2023; Esiobu et al. 2023). Still, further evaluations are needed to cover risks of harm, including addressing more nuanced gaps in risk areas for which evaluations exist.

Our third observation is that insofar as evaluation tools exist to address risks from generative AI, they are mainly evaluating model capabilities, with few studying human-AI interactions or the systemic impacts of such systems (Figure 3). This is parallel to gaps observed across machine learning domains (Hutchinson et al. 2022), reflecting a widely held assumption that the evaluation of these systems can be abstracted out of context (Selbst et al. 2019). More detailed inspection of the repository indicates that, among technical artifacts, safety evaluations focus particularly on AI system outputs and to a lesser degree on other artifacts such as training data, classifiers or filters. This emphasis on capabilities in isolation also results from how many evaluations have recently been performed and disclosed as part of large generative AI system announcements, which primarily focus on capability evaluations (Glaese et al. 2022; Anthropic 2023; OpenAI 2023b; Mishkin et al. 2022; Touvron et al. 2023b; Anil et al. 2023). It is also likely influenced by the inavailability of training data or filters used in proprietary models, contrasted with the availability of the AI system for black-box testing via public APIs.

While a capability-focused approach provides important indications as to potential downstream harms, and is therefore a core piece of safety evaluation, it does not account for contextual factors that co-determine risks of harm. Thus, it must be complemented by further analyses that add layers of relevant context. Without such context, it is only possible to measure what has already been anticipated, and it is challenging to create grounded thresholds for what good real world performance looks like.

Closing Safety Evaluation Gaps

Our assessment of the current state of safety evaluations of generative AI systems identified significant gaps in the modalities, risks and contexts covered by current evaluations. We now propose practical steps that can be taken to close those gaps, including a discussion of their advantages and limitations.

Increasing Modality Coverage

One gap in the current safety evaluation landscape is that tests for image-, audio-, video-, and multimodal AI systems are often missing. To increase modality coverage, there are some low-hanging fruit strategies - specifically, text-based evaluations can be **repurposed** to assess other modalities (e.g. image), or output in non-text modalities can be **transcribed** into text such that text-based evaluations can be applied. Here, we describe these approaches and their limitations.

Repurposing Evaluations One way to address gaps in the evaluation landscape is to repurpose components of existing evaluation methods. Repurposing and reusing datasets and tasks is a common approach in machine learning research, including on generative AI systems. For example, a sentiment bias evaluation that was introduced in 2019 by Huang et al. (2020) was cited in the GPT-3 paper for a modified sentiment bias analysis (Brown et al. 2020a). In their 2021 paper presenting Gopher, Rae et al. (2022) conducted the same analysis but used an expanded set of prompts. Most recently, PaLM2 drew on the Gopher prompt set for a multilingual toxicity analysis (Anil et al. 2023). This practice of reuse is especially common when AI system developers are working on tight timelines in fast-moving research domains, as is the case with generative AI.

However, this approach must be pursued with caution, as repurposing can create a range of negative unintended effects. For example, Winogender (Rudinger et al. 2018) and Winobias (Zhao et al. 2018) were developed as benchmarks to address the specific problem in language modelling of coreference resolution. These benchmarks are now commonly repurposed for the different goal of assessing bias in large generative AI systems, as they quantify the association of gender and occupation in text output. The basic framework of these evaluations was furthermore used as inspiration for probing risks in images, as seen in the DALL-E 2 system card (Mishkin et al. 2022). In this propagation, a narrow evaluation (as is appropriate for coreference resolution) has become state-of-the-art on an inherently broad and manifold problem, social discriminatory bias. In this narrow test, several forms of social bias go unmeasured. Interestingly, the narrow operationalisation of social bias in language models has even carried over to other modalities, where no prior evaluations are being repurposed - rather,

novel evaluations are being developed from scratch, equally operationalising social bias via the association of gender and occupation, e.g. to assess bias in image models (Luccioni et al. 2024; Naik and Nushi 2023). One step to mitigate this is to carefully capture and document the propagation of specific evaluations across different use cases (Koch et al. 2021; Bommasani et al. 2023).

Another way in which repurposing may occur is by relying on classifiers as evaluation tools. However, such repurposing is risky and may return distorted or invalid results. Specifically, in all forms of repurposing lies a risk that the evaluation in question may not be valid in the context to which it is being repurposed. We now survey these risks and ways to address them.

While repurposing saves work and can create common standards, applying an evaluation or classifier out of its intended context comes with trade-offs such that repurposing, if done poorly, may create more harm than good (Selbst et al. 2019; Rauh et al. 2022). For example, repurposing an evaluation may result in simply invalid results - a metric that is appropriate in one context or modality may be invalid in another context or modality. If a risk is being evaluated in this way, this may not create any helpful insights and, critically, provide a false sense of security.

To sidestep these risks, rather than directly repurposing existing evaluations to assess risks in other modalities, another option is to use existing tools as a starting point for refinement. By using existing evaluations as a guiding analogy for constructing new evaluations, researchers could adapt or augment prior evaluation datasets.

To determine if and when an evaluation should be reused, practitioners should consider its provenance, identify how the original context and purpose align with the new usage, and understand what norms are being perpetuated by its reuse. Because risks of harm are contextual, understanding the difference between the original and updated context will uncover the gaps in the new use case. This is especially important because some information inevitably gets lost when operationalising complex constructs (e.g. a certain societal risk) such that they can be measured – translating risks from AI systems into narrow metrics and tests is fraught with ambiguity (Wagner et al. 2021). This loss compromises the validity of a measure. Here we are particularly concerned with *external validity* failures, where a test may capture a target construct fully in a given instance but not allow extrapolation to new situations (Liao et al. 2021).

In addition to considering the differences between the evaluation's original and new context, repurposing an evaluation requires reflection on how it operationalises harm and how it may offer opportunities for refinement. Evaluations operationalise broad concepts, e.g. misinformation, into more specific measures, e.g. fact-checking accuracy on human-written COVID-19 information (Bang et al. 2023). Without this reflection and refinement, repurposing evaluations may lead to missing the actual intended evaluation target.

Transcribing non-text output for text-based evaluation Another way to address the uneven distribution of evaluations across modalities is to transcribe outputs from one modality into another modality in which an appropriate evaluation exists. This may be attempted through transcribing content from images, video, or audio output such that the transcript can then be evaluated using text-based evaluation tools. For example, automatic speech recognition tools can be leveraged to transcribe speech into text or an image captioning system can be used to caption a generated image (e.g. Wiles, Albuquerque, and Gowal (2023)). Similarly, video can be split into a series of images to enable image-based evaluation.

Transcription into text is a valuable and tractable first step in evaluating risks of harm in non-text modalities. However, through the process of transcription, some information inevitably gets lost and thus evades evaluation. For example, in speech, prosody (the way in which something is said, e.g. with sarcasm) carries information about meaning but this may not be captured by simple transcription of sentences (Wilson and Wharton 2006). Similarly, generating synthetic audio in the voice of a particular person may create appropriation or defamation harms that would not be detected by transcribing the audio and analysing the text. Additional pitfalls of the transcription approach stem from the fact that methods to translate between modalities may be error-prone (Rohrbach et al. 2018; Ramesh, KhudaBukhsh, and Kumar 2022), sometimes in systematically biased ways (Ngueajio and Washington 2022; Wang et al. 2022). Such errors can propagate through the harm analysis. For example, if an image-captioning system is biased toward people on motorcycles as "male" (Hendricks et al. 2018), evaluation of image captions may indicate a different gender bias than is present in the images that are the target of evaluation. In sum, while transcription approaches are a promising first step, these methods are limited, require quality checks, and must be complemented by evaluation methods that are calibrated to the output modality directly.

Increasing Risk Coverage

Model-driven or "automated" evaluations are one way to increase coverage across risk areas. We introduce these methods as immediately available low-hanging fruit. However, these approaches are limited and ultimately not sufficient.

Model-driven evaluations Generative AI systems themselves are being used as evaluation tools because of the adaptability they offer for coverage of risks. Language models can be used to adversarially generate model inputs that elicit harmful language from other language models (Perez et al. 2022a) or lead to systematic failures in model predictions (Wiles, Albuquerque, and Gowal 2023). In contrast to asking humans to generate model inputs, generating inputs automatically often requires little work in the form of prompting or fine-tuning on small amounts of data, compared to a potentially time-consuming human evaluation task. Language models have also been used to classify output text as harmful or not (Bai et al. 2022; Inan et al. 2023; Markov et al. 2023). Instead of relying on human annotators to manually label data and training bespoke classifiers for individual harms, language models enable quick implemen-

tation of classifiers with little to no additional data. Building evaluation tools with language models, as opposed to relying on human annotators, allows for flexible pipelines that are fast for researchers to iterate on. As such, they make it easy to adjust and expand definitions of harmful behaviour to changing model capabilities, emergent risks, or usage patterns. They can also mitigate the drawbacks of evaluations relying on human raters, which are typically costly and slow, and could put the raters themselves at risk when the evaluation task is disturbing. Although work in this space is currently limited to text, there may be ways to expand model-driven evaluations, such as adversarial probing, to other modalities.

However, AI systems as evaluation tools face limitations. Current work tends to rely on proprietary AI systems that may not be accessible to those performing an evaluation. It also limits trust in such evaluation results, as the system used cannot be inspected and methods may not be replicable. AI systems are also updated over time and produce samples stochastically, which may adversely impact the reproducibility of this approach (Pangakis, Wolken, and Fasching 2023). In addition, generative models may have biases and behave in unexpected ways, which can introduce confounds or noise to the evaluation. There is a further risk of spiralling effects if AI systems from the same base model are used to evaluate each other, as existing biases or blindspots present in these systems can be amplified through this process.

This method is also limited in the types of risks it can address: it is most readily useful for covering risks from "unsafe" outputs, rather than risks that manifest during human interaction or widespread deployment of an AI system. As described above, the model-driven evaluations currently being developed center on evaluation of the technical artefact in isolation and use the model as a classifier or automated probe. Finally, this direction of evaluation is novel, and its robustness needs to be assessed. Grounding the results of model-driven evaluations by comparing them with human or well-established evaluations is a cross-validation step to ensure this method does not fall foul of validation problems.

Increasing Context Coverage

Most safety evaluations to date focus on the model, and do not take into account human-AI-interaction or broader societal context in which an AI system may be deployed. Closing the context gap will require new paradigms for evaluation, but this does not require inventing new methods from scratch: rather, interdisciplinary groups can expand the toolbox to tried-and-tested methods from disciplines that have not traditionally been used in AI development, and apply those to AI safety evaluation.

Looking beyond 'model-centric' evaluation methods Our analysis of the evaluation landscape surfaced substantial context gaps, finding that most AI harms identified only reflect behaviour and outcomes of the model in isolation. This lack of coverage is closely related to an overrepresentation of certain methodologies, particularly automated benchmarks, which are familiar to and favored by AI practitioners. Relying on a handful of well-established tools limits the extent to which evaluation can uncover harms that manifest outside of the model-centered context. Closing context gaps in safety evaluation may require a paradigm shift based around a simple principle: evaluating the safety of an AI system in its proper context requires methods that can take that context into account.

Interdisciplinary approaches to safety evaluation Expanding the methodological toolkit for a sociotechnical approach to safety evaluation does not require methodological innovation. Rather, it can be done by embracing a more interdisciplinary approach to safety evaluation. Experimental methodologies from the discipline of human-computer interaction or the social sciences are particularly useful. They take into account broader system-level metrics, such as economic indicators, to predict or assess impacts on broader systems, such as the labour market. For example, longitudinal studies can inform the design of evaluations aimed at uncovering long-ranging societal impacts of AI systems, e.g. trends in occupational values in light of AI-induced automation (Långstedt, Spohr, and Hellström 2023) and patterns of behaviour in engagement with AI-enabled assistants (Xu and Li 2022).

Qualitative analyses of how people perceive, work with, and relate to AI systems are another rich information source for safety evaluation. There is a growing body of work investigating the effect of interactive AI systems, e.g. humans' ability to distinguish AI- and human-generated content (Jakesch, Hancock, and Naaman 2023; Waltzer et al. 2023) and how users display attachment to AI "companions" (Laestadius et al. 2022; Pentina, Hancock, and Xie 2023). Ethnographic assessments of the functionality and safety failures at the point of real-world use are a further set of approaches that can yield new insights on the safety of a system, e.g. Marda and Narayan (2020); Raji et al. (2022).

Human-computer interaction research, behavioural studies, user research, ethnographic studies and social, economic and environmental impact assessments all present established fields with validated frameworks, metrics, and measurement approaches that have been applied to evaluate the safety of generative and other types of AI systems in their proper contexts (e.g. Elish and Watkins (2020); Marda and Narayan (2020); Brynjolfsson, Li, and Raymond (2023); Peng et al. (2023); Noy and Zhang (2023)). These approaches can be leveraged more systematically to comprehensively assess the safety of generative AI systems in relevant contexts, such as specific use cases, user groups, or institutions in which AI systems may be deployed. We list an overview of current and prospective safety evaluation methods in Appendix A.1 (https://dpmd.ai/3WUc5un). Moreover, incorporating insights from diverse disciplines infuses safety evaluations with a richer set of perspectives. This opens doors to evaluating risks that might otherwise be overlooked and critically examining conclusions drawn from individual evaluation approaches.

To meaningfully integrate these methodologies into practice, the necessary resources and infrastructure must be provided to interdisciplinary researchers both within and beyond the immediate AI safety and ethics community. There

are many avenues to enabling critical interdisciplinary safety work. For large technology corporations, it may involve onboarding a range of expertise by hiring with an interdisciplinary focus. Complementary to this approach, organisations could seek to provide seed-funding for novel approaches to AI safety research (for an example, see Superalignment Fast Grants (OpenAI 2023). Conferences on AI safety can function as venues to bring different disciplines together and foster collaboration, including between different "safety communities" (Weidinger et al. 2024). Individual research teams and institutions alike may look to establish long-standing collaborations or partnerships with other expert organisations – including but not limited to those that specialise in AI evaluations, to share expertise and resources. Standing up new interdisciplinary collaborations for safety evaluation requires an upfront cost in terms of time, infrastructure and resources.

Limitations of Safety Evaluation

Evaluations are critical tools on the path to building safe generative AI systems. Different methods and modes of evaluation can be leveraged to build an increasingly fuller picture of the risks of harm that these models pose; thus supporting risk mitigation, accountability, and responsible decision-making by developers and users alike. However, evaluation is limited and on its own cannot guarantee safe AI systems. In this section, we discuss limitations of evaluation in detail. First, simply expanding the scope of safety evaluations is no guarantee of completeness. Rather, comprehensive safety evaluation requires continually assessing and adapting to the risk profiles of specific use cases offered by generative AI systems. This includes the need for continuous monitoring of the ways in which these systems are (mis)used in the real world, which may not always be easy to anticipate at first point of deployment. Second, even where the use case and risk profile are well-understood, evaluations are constrained by the choices of their creators. Evaluations are inherently shaped by the values and biases encoded in their design decisions, and are never neutral. Finally, evaluation is not enough to ensure sociotechnical safety of a generative AI system. To truly build safe AI systems, evaluation must be embedded into broader responsible innovation practices and safety ecosystems that act on evaluation results through mitigation and governance efforts.

Evaluating 'General-purpose' Systems

The challenges of evaluation are particularly apparent in the context of "general-purpose" generative AI systems, whose downstream application or user base is not yet defined or understood. An often-cited ambition in the innovation of generative AI systems is to develop "general-purpose technologies" that could be applied to a wide range of potential tasks and environments (e.g. Bubeck et al. (2023) though see also, (Raji et al. 2021)). Indeed, generative AI systems have been likened to general-purpose technologies such as steam engines and office automation (Acemoglu and Johnson 2023). This supposed open-endedness of AI systems can make it difficult to identify the contexts – such as applications, user groups, or institutions – in which AI system safety should be evaluated.

One way to address this tension in practice is to define hypothetical applications of "general-purpose technologies" and to evaluate them in these contexts. This can, for example, take the form of identifying "critical user journeys", i.e. mapping a series of steps users may take while using a product to achieve a desired outcome (Arguelles et al. 2020). Following a precautionary approach, such hypothetical use-case mapping may first focus on high-risk applications. Still, early evaluation based on hypothetical use cases cannot replace downstream evaluation of actual use cases; rather, it serves to highlight risks and must be complemented with monitoring of real-world impacts. The risk profiles and thresholds of what constitutes "acceptable" model performance may differ between different downstream applications or user groups, requiring more rigorous evaluation in some cases than in others.

Rather than evaluating "general-purpose" AI systems as such, much work is needed to identify and evaluate risks in the context of their diverse potential use cases. Correspondingly, safety evaluations are not general-purpose either: rather, evaluations need to fit the specific risk areas, modalities, and contexts that they are intended to address. Evaluation requires explicitly accounting for the variability of factors that co-determine risk (e.g. evaluating the risk for different use cases). In this way, evaluation is specific and grounded, just like a use case of an AI system is specific and grounded - and neither AI systems nor safety evaluation are general-purpose solutions.

Evaluation Is Never Value-neutral

Evaluation necessarily and inherently covers only a subset of all possible manifestations of risks of harm (Bergman et al. 2023; Raji et al. 2021). What is included depends on pragmatic and normative considerations, such as what is tractable, anticipated, and prioritised. Thus, designing an evaluation involves choices – made either deliberately or implicitly – on what to prioritise and what to discard. Even the very first step of selecting a target to evaluate requires a normative judgement on what harms are important or relevant to measure (Kalluri 2020; Mohamed, Png, and Isaac 2020). Indeed, as the field makes progress on addressing the evaluation gaps we identified, the prioritisation of combinations of modality, risk, and context will be a value-driven process. Widespread adoption of such evaluations will then in turn perpetuate the values they encode (Bommasani 2023).

Furthermore, operationalising a harm construct into a metric necessarily bakes in certain assumptions. Operationalising the harm requires normative decisions on what task is most valuable for the system to perform highly on, how performance is measured, and where or to whom it is most valuable for the benefits of the system to accrue. After this process, what remains within scope of the evaluation is what is prioritised, and these decisions inherently express value. For example, making a commitment to a test and a metric – e.g. that social biases can be measured via associations of gender and occupations – is a normative judgement on where harms are likely to occur and which particular as-

pects of a harm are relevant and tractable (Luccioni et al. 2024). The way in which values are encoded into an evaluation depends on the corresponding risk, harm, and modality an evaluation aims to measure. For example, some values, such as those concerning beauty standards and body image, may not always be present in text, but are immediately visible in images and video. These normative decisions are all the more significant, as they tend to have a sticking effect that propagates as evaluations are being repurposed (see section "Results: Safety evaluation gaps"). Thus, when building evaluations in new modalities, we urge practitioners to carefully question the values underlying prior metrics.

Depending on what risk is being evaluated and how harm is operationalised, conducting an evaluation may even be inappropriate or problematic, or may create a disproportionate burden on those at risk of harm. For example, measuring sensitive traits to assess performance across demographic groups may place communities at risk or sit in tension with privacy, respect, or dignity (e.g. Wenger et al. (2022); Wolff (2010)). Characteristics or qualities that are essentially contested or fundamentally fluid (e.g. ethnicity, sexual orientation, or gender identity) may be reified through evaluations that bin these into discrete categories (Keyes 2019; Lu, Kay, and McKee 2022; Tomasev et al. 2021; Hanna et al. 2020). Some communities may not even wish to be represented in the evaluation (Denton et al. 2021; MediaWell 2019), either due to the burden (e.g. time and labour costs) of participation or, for example, if inclusion within the scope of the evaluation means being surveilled (Brunton and Nissenbaum 2016; Keyes 2019; MediaWell 2019; Bedoya 2014; Hassein 2017).

These normative decisions show that evaluations are inherently value expressions of those who conduct them. Therefore, the inclusion of perspectives beyond those who develop and deploy these systems is paramount. For example, assessing whether a model meets expectations prior to or after deployment requires a normative evaluation of whether some measurement expresses performance that is "good", "bad", "safe enough", etc. (see Bakalar et al. (2021)). However, for such thresholds to be legitimate, they need to arise from adequate institutions or processes, such as expert groups, democratic institutions, or fair and inclusive deliberation processes that centre groups that may be affected by these AI systems. Calls for greater representation of community groups is widespread and often offered as a mitigation for a broad range of fairness and sociotechnical harms (e.g. Costanza-Chock (2020); DeVries et al. (2019); National Institute of Standards and Technology (2021a); Sortition Foundation (2023); Lashbrook (2018); Jindal (2021); Pasquale and Malgieri (2021); Suresh and Guttag (2021); European Commission (2021)). While greater inclusion of groups and perspectives in the evaluation can lead to better visibility of the model performance (Buolamwini and Gebru 2018), there are arguments for adopting this approach with care (Bergman et al. 2023) due to the risk of disproportionate burden discussed above, as well as the potential for an oversimplified interpretation of the notion of representation and of the implementation of these approaches, which in turn may lead to objectification and exploitation (e.g. Fussell (2019)).

Evaluation Is Not All You Need

Evaluation is a cornerstone of responsible innovation: by transforming foresight and observed incidents into actionable insights, it underpins mitigation and responsible decision-making. While expanding sociotechnical evaluations is essential to address the gaps in the generative AI safety landscape, we also argue that it is important to recognize the limits of what evaluation can provide. Evaluation alone is not a panacea: to ensure the safe development of AI systems, it must be embedded in a broader responsible approach.

Even with the most rigorous evaluation process, certain types of harms will inevitably remain undetected, especially when deploying flexible and broadly applicable AI systems. This is why evaluation must be complemented with effective governance mechanisms. These should include pre-deployment assessments to evaluate remaining uncertainties, continuous post-deployment monitoring (including logging observed incidents (AI Incident Database 2023)), and swift, effective recourse mechanisms for those who experience or detect harm. Crucially, AI systems must be designed with flexibility, allowing new insights to be quickly translated into fixes, such as system updates. Given the pre-deployment evaluation gaps, organisations deploying AI systems require adequate governance infrastructures that can respond to detected risks with mitigations, either by delaying or stopping the deployment of an AI system or by suspending an already-deployed system until concerns are resolved. Only by embedding evaluation into meaningful processes can it lead to action and have actual impact on safety.

Conclusion

Generative AI systems present considerable safety risks, which can be holistically evaluated by adopting a sociotechnical approach. In this work, we survey the state of safety evaluations of generative AI. We find that the current safety evaluation landscape presents three significant gaps. Non-text modalities are under-served, especially for audio and video. Certain risk areas are also poorly covered in current evaluations. Finally, there are still too few evaluations that consider AI system safety beyond the scope of model outputs.

These gaps present a roadmap for the field of safety evaluation research. We release the repository (https://dpmd.ai/EvalsRepo) for the community to further analyse and use as a resource in their own work. Progress on closing these gaps will come from a combination of both short and long term work, extending present-day evaluation approaches as well as researching new paradigms. Although evaluation of "general-purpose" systems is challenging and will perpetuate whichever values are encoded in their design, it forms a critical component of responsible AI development.

Ethical Statement

Ethical Consideration Statement

This paper did not involve experiments with users or deployed systems, nor did it rely on sensitive user data.

As we conducted the review of evaluations to include in the repository, we were mindful of several ethical concerns, including how our backgrounds might introduce biases or a lack of objectivity into the review process. We addressed this by establishing a clear understanding within our team of the scope of this survey, including outlining its limitations, and devising clear criteria for inclusion. We also underwent multiple rounds of peer review to critically evaluate our work and methodology.

Researcher Positionality Statement

Our research team comprises research scientists and engineers from various national backgrounds within Western advanced industrialized countries, with the majority of us born in Europe or in the United States. Our operating language is English. Our collective experiences, cultural contexts and educational backgrounds have likely instilled particular methodologies and ways of parsing literature sources, which may have shaped our understanding of the field of study and, therefore, the scope of our review.

Adverse Impact Statement

We acknowledge that the mapping we provide in this paper is time-bound and necessarily limited. Any inadvertent omission of evaluations in our repository may provide a skewed or partial understanding of the state of the art. For that reason, we see it as critical for other teams with more diverse perspectives and epistemological traditions to undertake similar efforts to ensure that our collective understanding of the field of safety evaluations is balanced and does not overlook crucial insights.

Furthermore, we note that we did not perform quality control on the evaluations included in the repository (including ethical assessments of human data collection practices). This may have adverse impact if these evaluations were applied indiscriminately and without proper assessment of their limitations. Inclusion in the repository should not be seen as an endorsement of a given piece of work; rather it is a snapshot of what is being done to date.

Acknowledgements

We thank Simon Osindero, Sasha Brown, Matt Botvinick, Suresh Venkatasubramanian, Victor Ojewale, Fernando Diaz, Olivia Wiles, Doug Fritz, Courtney Biles, Nicklas Lundblad, Neil Rabinowitz, Jenny Brennan, Sunipa Dev, Don Wallace, Mark Díaz, Michal Lahav, Alex Kaskasoli, Isabela Albuquerque, Seliem El-Sayed, and Rida Qadri for their feedback and contributions to this work.

References

Acemoglu, D.; and Johnson, S. 2023. *Power and progress: our thousand-year struggle over technology and prosperity*. New York: PublicAffairs, first edition edition. ISBN 9781541702530.

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023a. MusicLM: Generating Music From Text. ArXiv:2301.11325 [cs, eess].

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023b. MusicLM: Generating Music From Text. arXiv:2301.11325.

AI Incident Database. 2023. Welcome to the Artificial Intelligence Incident Database. https://incidentdatabase.ai/.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J. L.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Bińkowski, M. a.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, 23716–23736. Curran Associates, Inc.

Alba, D. 2023. How Fake AI Photo of a Pentagon Blast Went Viral and Briefly Spooked Stocks. *Bloomberg.com*.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Étienne Goffinet; Hesslow, D.; Launay, J.; Malartic, Q.; Mazzotta, D.; Noune, B.; Pannier, B.; and Penedo, G. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867.

Amoroso, R.; Morelli, D.; Cornia, M.; Baraldi, L.; Del Bimbo, A.; and Cucchiara, R. 2023. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. ArXiv:2304.00500 [cs].

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; Chu, E.; Clark, J. H.; Shafey, L. E.; Huang, Y.; Meier-Hellstern, K.; Mishra, G.; Moreira, E.; Omernick, M.; Robinson, K.; Ruder, S.; Tay, Y.; Xiao, K.; Xu, Y.; Zhang, Y.; Abrego, G. H.; Ahn, J.; Austin, J.; Barham, P.; Botha, J.; Bradbury, J.; Brahma, S.; Brooks, K.; Catasta, M.; Cheng, Y.; Cherry, C.; Choquette-Choo, C. A.; Chowdhery, A.; Crepy, C.; Dave, S.; Dehghani, M.; Dev, S.; Devlin, J.; Díaz, M.; Du, N.; Dyer, E.; Feinberg, V.; Feng, F.; Fienber, V.; Freitag, M.; Garcia, X.; Gehrmann, S.; Gonzalez, L.; Gur-Ari, G.; Hand, S.; Hashemi, H.; Hou, L.; Howland, J.; Hu, A.; Hui, J.; Hurwitz, J.; Isard, M.; Ittycheriah, A.; Jagielski, M.; Jia, W.; Kenealy, K.; Krikun, M.; Kudugunta, S.; Lan, C.; Lee, K.; Lee, B.; Li, E.; Li, M.; Li, W.; Li, Y.; Li, J.; Lim, H.; Lin, H.; Liu, Z.; Liu, F.; Maggioni, M.; Mahendru, A.; Maynez, J.; Misra, V.; Moussalem, M.; Nado, Z.; Nham, J.; Ni, E.; Nystrom, A.; Parrish, A.; Pellat, M.; Polacek, M.; Polozov, A.; Pope, R.; Qiao, S.; Reif, E.; Richter, B.; Riley, P.; Ros, A. C.; Roy, A.; Saeta, B.; Samuel, R.; Shelby, R.; Slone, A.; Smilkov, D.; So, D. R.; Sohn, D.; Tokumine, S.; Valter, D.; Vasudevan, V.; Vodrahalli, K.; Wang, X.; Wang, P.; Wang, Z.; Wang, T.; Wieting, J.; Wu, Y.; Xu, K.; Xu, Y.; Xue, L.; Yin, P.; Yu, J.; Zhang, Q.; Zheng, S.; Zheng, C.; Zhou, W.; Zhou, D.; Petrov, S.; and Wu, Y. 2023. PaLM 2 Technical Report. ArXiv:2305.10403 [cs].

Anthropic. 2023a. Challenges in evaluating AI systems. https://www.anthropic.com/news/evaluating-ai-systems.

Anthropic. 2023b. Core Views on AI Safety: When, Why, What, and How. https://www.anthropic.com/index/core-views-on-ai-safety.

Anthropic. 2023. Model Card and Evaluations for Claude Models. Technical report, Anthropic.

Arguelles, C.; Sampson, T.; Kubik, J.; and Bibi, E. 2020. Critical User Journey Test Coverage. *Defensive Publications Series*.

Ashraf, N.; Rafiq, A.; Butt, S.; Shehzad, H. M. F.; Sidorov, G.; and Gelbukh, A. 2022. YouTube based religious hate speech and extremism detection dataset with machine learning baselines. *Journal of Intelligent & Fuzzy Systems*, 42(5): 4769–4777.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. ArXiv:2212.08073 [cs].

Bakalar, C.; Barreto, R.; Bergman, S.; Bogen, M.; Chern, B.; Corbett-Davies, S.; Hall, M.; Kloumann, I.; Lam, M.; Candela, J. Q.; Raghavan, M.; Simons, J.; Tannen, J.; Tong, E.; Vredenburgh, K.; and Zhao, J. 2021. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems. ArXiv:2103.06172 [cs].

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023.

Barnett, J. 2023. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 146–161. ArXiv:2307.05527 [cs, eess].

Bedoya, A. M. 2014. Big Data and the Underground Railroad. *Slate*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Bergman, A. S.; Hendricks, L. A.; Rauh, M.; Wu, B.; Agnew, W.; Kunesch, M.; Duan, I.; Gabriel, I.; and Isaac, W. 2023. Representation in AI Evaluations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 519–533. Chicago IL USA: ACM. ISBN 9798400701924.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. ArXiv:2211.03759 [cs].

Bird, C.; Ungless, E. L.; and Kasirzadeh, A. 2023. Typology of Risks of Generative Text-to-Image Models. ArXiv:2307.05543 [cs].

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. ArXiv:2110.01963 [cs].

Bommasani, R. 2023. Evaluation for Change. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8227–8239. Toronto, Canada: Association for Computational Linguistics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 [cs].

Bommasani, R.; Soylu, D.; Liao, T. I.; Creel, K. A.; and Liang, P. 2023. Ecosystem Graphs: The Social Footprint of Foundation Models. arXiv:2303.15772.

Borchers, C.; Gala, D.; Gilburt, B.; Oravkin, E.; Bounsi, W.; Asano, Y. M.; and Kirk, H. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 212–224. Seattle, Washington: Association for Computational Linguistics.

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; and Zeghidour, N. 2023. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2523–2533.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell,

A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020a. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bruell, A. 2023. BuzzFeed to Use ChatGPT Creator OpenAI to Help Create Quizzes and Other Content. *Wall Street Journal*.

Brunton, F.; and Nissenbaum, H. 2016. *Obfuscation: a user's guide for privacy and protest*. Cambridge, Mass. London: MIT Press. ISBN 9780262529860.

Brynjolfsson, E.; Li, D.; and Raymond, L. R. 2023. Generative AI at Work. Working Paper 31161, National Bureau of Economic Research.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramèr, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023a. Extracting Training Data from Diffusion Models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 5253–5270. Anaheim, CA: USENIX Association. ISBN 978-1-939133-37-3.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Awadalla, A.; Koh, P. W.; Ippolito, D.; Lee, K.; Tramer, F.; and Schmidt, L. 2023b. Are aligned neural networks adversarially aligned? ArXiv:2306.15447 [cs].

Cho, J.; Zala, A.; and Bansal, M. 2023. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. ArXiv:2202.04053 [cs].

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Cohere Model Limitations. 2023. https://docs.cohere.com/docs/model-limitations.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation. arXiv:2306.05284.

Costa-jussà, M.; Andrews, P.; Smith, E.; Hansanti, P.; Ropers, C.; Kalbassi, E.; Gao, C.; Licht, D.; and Wood, C. 2023. Multilingual Holistic Bias: Extending Descriptors and Patterns to Unveil Demographic Biases in Languages at Scale. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14141–14156. Singapore: Association for Computational Linguistics.

Costanza-Chock, S. 2020. *Design justice: community-led practices to build the worlds we need*. Information policy. Cambridge, Massachusetts: MIT Press. ISBN 9780262043458.

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1571–1583. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Das, M.; Raj, R.; Saha, P.; Mathew, B.; Gupta, M.; and Mukherjee, A. 2023. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17: 1014–1023.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 205395172110359.

DeVries, T.; Misra, I.; Wang, C.; and van der Maaten, L. 2019. Does Object Recognition Work for Everyone? Technical report, Meta.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A Generative Model for Music. ArXiv:2005.00341 [cs, eess, stat].

Dinan, E.; Abercrombie, G.; Bergman, A. S.; Spruit, S.; Hovy, D.; Boureau, Y.-L.; and Rieser, V. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. ArXiv:2107.03451 [cs].

Du, Y.; Chen, Z.; Salamon, J.; Russell, B.; and Owens, A. 2023. Conditional Generation of Audio from Video via Foley Analogies. ArXiv:2304.08490 [cs, eess].

Electronic Privacy Information Center. 2023. Regarding the Artificial Intelligence Risk Management Framework. https://epic.org/documents/regarding-the-artificial-intelligence-risk-management-framework/.

Elish, M. C.; and Watkins, E. A. 2020. Repairing Innovation: A Study of Integrating AI in Clinical Care. Technical report, Data & Society.

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. ArXiv:2303.10130 [cs, econ, q-fin].

Esiobu, D.; Tan, X.; Hosseini, S.; Ung, M.; Zhang, Y.; Fernandes, J.; Dwivedi-Yu, J.; Presani, E.; Williams, A.; and Smith, E. 2023. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3764–3814. Singapore: Association for Computational Linguistics.

EU AI Act. 2023. Regulatory framework proposal on artificial intelligence: Shaping Europe's digital future.

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

Ferrara, E. 2024. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*.

Fussell, S. 2019. How an Attempt at Correcting Bias in Tech Goes Wrong. *The Atlantic*.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; El-Sayed, S.; Brown, S.; Akbulut, C.; Trask, A.; Hughes, E.; Bergman, A. S.; Shelby, R.; Marchal, N.; Griffin, C.; Mateos-Garcia, J.; Weidinger, L.; Street, W.; Lange, B.; Ingerman, A.; Lentz, A.; Enger, R.; Barakat, A.; Krakovna, V.; Siy, J. O.; Kurth-Nelson, Z.; McCroskery, A.; Bolina, V.; Law, H.; Shanahan, M.; Alberts, L.; Balle, B.; de Haas, S.; Ibitoye, Y.; Dafoe, A.; Goldberg, B.; Krier, S.; Reese, A.; Witherspoon, S.; Hawkins, W.; Rauh, M.; Wallace, D.; Franklin, M.; Goldstein, J. A.; Lehman, J.; Klenk, M.; Vallor, S.; Biles, C.; Ringel Morris, M.; King, H.; Agüera y Arcas, B.; Isaac, W.; and Manyika, J. 2024. The Ethics of Advanced AI Assistants. *preprint*.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; Campbell-Gillingham, L.; Uesato, J.; Huang, P.-S.; Comanescu, R.; Yang, F.; See, A.; Dathathri, S.; Greig, R.; Chen, C.; Fritz, D.; Elias, J. S.; Green, R.; Mokrá, S.; Fernando, N.; Wu, B.; Foley, R.; Young, S.; Gabriel, I.; Isaac, W.; Mellor, J.; Hassabis, D.; Kavukcuoglu, K.; Hendricks, L. A.; and Irving, G. 2022. Improving alignment of dialogue agents via targeted human judgements. ArXiv:2209.14375 [cs].

GPT-4V(ision) system card. 2023. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card.

Griffith, E. 2023. My Weekend With an Emotional Support A.I. Companion. *The New York Times*.

Hameleers, M.; Powell, T. E.; Van Der Meer, T. G.; and Bos, L. 2020. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, 37(2): 281–301.

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 501–512. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Hassein, N. 2017. Against Black Inclusion in Facial Recognition. *Digital Talking Drum*.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, 793–811. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-01218-2.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, P.-S.; Zhang, H.; Jiang, R.; Stanforth, R.; Welbl, J.; Rae, J.; Maini, V.; Yogatama, D.; and Kohli, P. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 65–83. Online: Association for Computational Linguistics.

Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; Engel, J.; Le, Q. V.; Chan, W.; Chen, Z.; and Han, W. 2023. Noise2Music: Text-conditioned Music Generation with Diffusion Models. ArXiv:2302.03917 [cs, eess].

Huang, S.; Grady, P.; and GPT-3. 2022. Generative AI: A Creative New World. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/.

Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation Gaps in Machine Learning Practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1859–1876. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Hutiri, W.; Papakyriakopoulos, O.; and Xiang, A. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 359–376. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674.

Inflection-1. 2023. Inflection-1 Tech Memo. https://inflection.ai/assets/Inflection-1_0622.pdf.

Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120.

Jindal, S. 2021. Responsible Sourcing of Data Enrichment Services. *Partnership on AI*.

Kalluri, P. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815): 169–169.

Keyes, O. 2019. Counting the Countless. *Real Life*.

Khlaaf, H.; Mishkin, P.; Achiam, J.; Krueger, G.; and Brundage, M. 2022. A Hazard Analysis Framework for Code Synthesis Large Language Models. ArXiv:2207.14157 [cs].

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. ArXiv:2005.04790 [cs].

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. arXiv:2303.05453.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. ArXiv:2112.01716 [cs, stat].

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 14614448221142007.

Långstedt, J.; Spohr, J.; and Hellström, M. 2023. Are our values becoming more fit for artificial intelligence society? A longitudinal study of occupational values and occupational susceptibility to technological substitution. *Technology in Society*, 72: 102205.

Lashbrook, A. 2018. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. *The Atlantic*.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; and Hsu, W.-N. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. arXiv:2306.15687.

Liao, T.; Taori, R.; Raji, D.; and Schmidt, L. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. ArXiv:2308.05374 [cs].

Lu, C.; Kay, J.; and McKee, K. 2022. Subverting machines, fluctuating identities: Re-learning human categorization. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1005–1015. Seoul Republic of Korea: ACM. ISBN 9781450393522.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable bias: evaluating societal representations in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Luccioni, A. S.; Viguier, S.; and Ligozat, A.-L. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. ADS Bibcode: 2022arXiv221102001S.

Malik, M.; and Johansson, R. 2022. Controlling for Stereotypes in Multimodal Language Model Evaluation. In Bastings, J.; Belinkov, Y.; Elazar, Y.; Hupkes, D.; Saphra, N.; and Wiegreffe, S., eds., *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 263–271. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Mandal, A.; Leavy, S.; and Little, S. 2023. Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models. ArXiv:2304.13855 [cs].

Marda, V.; and Narayan, S. 2020. Data in New Delhi's predictive policing system. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 317–324. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Markov, T.; Zhang, C.; Agarwal, S.; Nekoul, F. E.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

MediaWell. 2019. "Please do not include us": Workshop on AI Ethics and Inclusion – MediaWell.

Mishkin, P.; Ahmad, L.; Brundage, M.; Krueger, G.; and Sastry, G. 2022. DALL·E 2 Preview – Risks and Limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4): 659–684.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*.

Naik, R.; and Nushi, B. 2023. Social Biases through the Text-to-Image Generation Lens. ArXiv:2304.06034 [cs].

National Institute of Standards and Technology. 2021a. AI Risk Management Framework. *NIST*.

National Institute of Standards and Technology. 2021b. AI RMF Playbook. Technical report, National Institute of Standards and Technology, Gaithersburg, MD.

Nelson, D. L.; Reed, V. S.; and Walling, J. R. 1976. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5): 523–528.

Ngueajio, M. K.; and Washington, G. 2022. Hey ASR System! Why Aren't You More Inclusive? In Chen, J. Y. C.; Fragomeni, G.; Degen, H.; and Ntoa, S., eds., *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, 421–440. Cham: Springer Nature Switzerland. ISBN 978-3-031-21707-4.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of GPT-4 on Medical Challenge Problems. ArXiv:2303.13375 [cs].

Noy, S.; and Zhang, W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654): 187–192.

OECD Expert Group on AI Incidents. 2023. Expert Group on AI Incidents - OECD.AI.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. ArXiv:1609.03499 [cs] version: 2.

OpenAI. 2023a. DALL·E 3 system card. https://openai.com/research/dall-e-3-system-card.

OpenAI. 2023b. GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

OpenAI. 2023c. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

OpenAI. 2023d. Our approach to AI safety. https://openai.com/blog/our-approach-to-ai-safety.

OpenAI. 2023. Superalignment Fast Grants. https://openai.com/blog/superalignment-fast-grants.

Pangakis, N.; Wolken, S.; and Fasching, N. 2023. Automated Annotation with Generative AI Requires Validation. arXiv:2306.00176.

Pasquale, F.; and Malgieri, G. 2021. Opinion | If You Don't Trust A.I. Yet, You're Not Wrong. *The New York Times*.

Peng, S.; Kalliamvakou, E.; Cihon, P.; and Demirer, M. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. ArXiv:2302.06590 [cs].

Pentina, I.; Hancock, T.; and Xie, T. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140: 107600.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022a. Red Teaming Language Models with Language Models. ArXiv:2202.03286 [cs].

Perez, E.; Ringer, S.; Lukošiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A.; Chen, A.; Mann, B.; Israel, B.; Seethor, B.; McKinnon, C.; Olah, C.; Yan, D.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Khundadze, G.; Kernion, J.; Landis, J.; Kerr, J.; Mueller, J.; Hyun, J.; Landau, J.; Ndousse, K.; Goldberg, L.; Lovitt, L.; Lucas, M.; Sellitto, M.; Zhang, M.; Kingsland, N.; Elhage, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Larson, R.; McCandlish, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Lanham, T.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Clark, J.; Bowman, S. R.; Askell, A.; Grosse, R.; Hernandez, D.; Ganguli, D.; Hubinger, E.; Schiefer, N.; and Kaplan, J. 2022b. Discovering Language Model Behaviors with Model-Written Evaluations. ArXiv:2212.09251 [cs].

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517. Chicago IL USA: ACM. ISBN 9798400701924.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; Rutherford, E.; Hennigan, T.; Menick, J.; Cassirer, A.; Powell, R.; Driessche, G. v. d.; Hendricks, L. A.; Rauh, M.; Huang, P.-S.; Glaese, A.; Welbl, J.; Dathathri, S.; Huang, S.; Uesato, J.; Mellor, J.; Higgins, I.; Creswell, A.; McAleese, N.; Wu, A.; Elsen, E.; Jayakumar, S.; Buchatskaya, E.; Budden, D.; Sutherland, E.; Simonyan, K.; Paganini, M.; Sifre, L.; Martens, L.; Li, X. L.; Kuncoro, A.; Nematzadeh, A.; Gribovskaya, E.; Donato, D.; Lazaridou, A.; Mensch, A.; Lespiau, J.-B.; Tsimpoukelli, M.; Grigorev, N.; Fritz, D.; Sottiaux, T.; Pajarskas, M.; Pohlen, T.; Gong, Z.; Toyama, D.; d'Autume, C. d. M.; Li, Y.; Terzi, T.; Mikulik, V.; Babuschkin, I.; Clark, A.; Casas, D. d. L.; Guy, A.; Jones, C.; Bradbury, J.; Johnson, M.; Hechtman, B.; Weidinger, L.; Gabriel, I.; Isaac, W.; Lockhart, E.; Osindero, S.; Rimell, L.; Dyer, C.; Vinyals, O.; Ayoub, K.; Stanway, J.; Bennett, L.; Hassabis, D.; Kavukcuoglu, K.; and Irving, G. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. ArXiv:2112.11446 [cs].

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. ArXiv:2111.15366 [cs].

Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 959–972. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 33–44. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. ArXiv:2102.12092 [cs].

Ramesh, K.; KhudaBukhsh, A. R.; and Kumar, S. 2022. 'Beach' to 'Bitch': Inadvertent Unsafe Transcription of Kids' Content on YouTube. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12108–12118.

Rauh, M.; Mellor, J.; Uesato, J.; Huang, P.-S.; Welbl, J.; Weidinger, L.; Dathathri, S.; Glaese, A.; Irving, G.; Gabriel, I.; Isaac, W.; and Hendricks, L. A. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. *Advances in Neural Information Processing Systems*, 35: 24720–24739.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. Brussels, Belgium: Association for Computational Linguistics.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. ArXiv:2112.10752 [cs].

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 36479–36494. Curran Associates, Inc.

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

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. ArXiv:2210.05791 [cs].

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model evaluation for extreme risks. ArXiv:2305.15324 [cs].

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; Arcas, B. A. y.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. ArXiv:2305.09617 [cs].

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Daumé III, H.; Dodge, J.; Evans, E.; Hooker, S.; Jernite, Y.; Luccioni, A. S.; Lusoli, A.; Mitchell, M.; Newman, J.; Png, M.-T.; Strait, A.; and Vassilev, A. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. ArXiv:2306.05949 [cs].

Sortition Foundation. 2023. Sortition Foundation. https://www.sortitionfoundation.org/.

Stilgoe, J.; Owen, R.; and Macnaghten, P. 2013. Developing a framework for responsible innovation. *Research Policy*, 42(9): 1568–1580.

Stoev, T.; Yordanova, K.; and Tonkin, E. L. 2023. Experiencing Annotation: Emotion, Motivation and Bias in Annotation Tasks. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 534–539. Atlanta, GA, USA: IEEE. ISBN 9781665453813.

Sun, L.; Wei, M.; Sun, Y.; Suh, Y. J.; Shen, L.; and Yang, S. 2023. Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI. arXiv:2305.10566.

Suresh, H.; and Guttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, 1–9. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385534.

Team, G. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

The Vicuna Team. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/.

The White House. 2023. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.

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

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

UK Task Force. 2023. Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI. https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai. Accessed: 2023-09-28.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vincent, J. 2023. Republicans respond to Biden reelection announcement with AI-generated attack ad. *The Verge*.

Wagner, C.; Strohmaier, M.; Olteanu, A.; Kıcıman, E.; Contractor, N.; and Eliassi-Rad, T. 2021. Measuring algorithmically infused societies. *Nature*, 595(7866): 197–204.

Waltzer, T.; Cox, R. L.; Heyman, G. D.; et al. 2023. Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students. *Human Behavior and Emerging Technologies*, 2023.

Wang, A.; Barocas, S.; Laird, K.; and Wallach, H. 2022. Measuring Representational Harms in Image Captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 324–335. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Weidinger, L.; Barnhart, J.; Brennan, J.; Butterfield, C.; Young, S.; Hawkins, W.; Hendricks, L. A.; Comanescu, R.; Chang, O.; Rodriguez, M.; andDawn Bloxwich, J. B.; Proleev, L.; Chen, J.; Farquhar, S.; Ho, L.; Gabriel, I.; Dafoe, A.; and Isaac, W. 2024. Holistic Safety and Responsibility Evaluationsof Advanced AI Models. *arxiv*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and social risks of harm from Language Models. ArXiv:2112.04359 [cs].

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems.

Wenger, E.; Bhattacharjee, R.; Bhagoji, A. N.; Passananti, J.; Andere, E.; Zheng, H.; and Zhao, B. 2022. Finding Naturally Occurring Physical Backdoors in Image Datasets. *Advances in Neural Information Processing Systems*, 35: 22103–22116.

Wiles, O.; Albuquerque, I.; and Gowal, S. 2023. Discovering Bugs in Vision Models using Off-the-shelf Image Generation and Captioning. ArXiv:2208.08831 [cs, stat].

Wilson, D.; and Wharton, T. 2006. Relevance and prosody. *Journal of Pragmatics*, 38(10): 1559–1579.

Wolff, J. 2010. Fairness, Respect and the Egalitarian "Ethos" Revisited. *The Journal of Ethics*, 14(3/4): 335–350.

Wu, C. S.; and Bhandary, U. 2020. Detection of Hate Speech in Videos Using Machine Learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 585–590.

Xiang, C. 2023. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.

Xu, S.; and Li, W. 2022. A tool or a social being? A dynamic longitudinal investigation of functional use and relational use of AI voice assistants. *New Media & Society*, 14614448221108112.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. arXiv:2206.10789.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. ArXiv:1804.06876 [cs].