Sociotechnical Safety Evaluation of Generative AI Systems

Laura Weidinger[1], **Maribeth Rauh**[1], **Nahema Marchal**[1], **Arianna Manzini**[1], **Lisa Anne Hendricks**[1], **Juan Mateos-Garcia**[1], **Stevie Bergman**[1], **Jackie Kay**[1], **Conor Griffin**[1], **Ben Bariach**[1], **Iason Gabriel**[1], **Verena Rieser**[1] **and William Isaac**[1]

[1]Google DeepMind, London N1C 4DN, United Kingdom

Generative AI systems produce a range of risks. To ensure the safety of generative AI systems, these risks must be evaluated. In this paper, we make two main contributions toward establishing such evaluations. First, we propose a three-layered framework that takes a structured, sociotechnical approach to evaluating these risks. This framework encompasses capability evaluations, which are the main current approach to safety evaluation. It then reaches further by building on system safety principles, particularly the insight that context determines whether a given capability may cause harm. To account for relevant context, our framework adds human interaction and systemic impacts as additional layers of evaluation. Second, we survey the current state of safety evaluation of generative AI systems and create a repository of existing evaluations. Three salient evaluation gaps emerge from this analysis. We propose ways forward to closing these gaps, outlining practical steps as well as roles and responsibilities for different actors. Sociotechnical safety evaluation is a tractable approach to the robust and comprehensive safety evaluation of generative AI systems.

Keywords: Evaluation, Sociotechnical, Generative AI, Multimodal

Corresponding author(s): Laura Weidinger <lweidinger@deepmind.com>

Acknowledgements

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

Contents

Reader's guide

This is a long document. Depending on your background and interests, we recommend different reading strategies:

- **Two-minute read:** Look at figure 2.1 (p.10) that illustrates our three-layered evaluation framework, and figures 3.1-3 (p.15) which depict the current state of safety evaluations.
- **Ten-minute read:** Read the abstract and skim section 2 (p.7), which introduces our three-layered evaluation framework; look at figures 3.1-3 (p.15) which depict the current state of safety evaluations.
- **Evaluators:** Skim section 2 (p.7), where we introduce our three-layered evaluation framework, and section 3 (p.12) where we survey the current state of safety evaluation; dedicate most time to section 4 (p.17) on practical steps to closing evaluation gaps, and to the case study (p.39) on evaluating misinformation that puts our evaluation framework into practice. Read about evaluation as a practice of responsible innovation in the discussion (p.22), and about the limitations of specific evaluation methodologies (p.32).
- **People steering AI labs:** Look at figure 2.1 (p.10) that illustrates our three-layered evaluation framework, read section 3 (p.12), which outlines gaps in the current state of safety evaluation of generative AI systems, and look at figure 5 (p.23) that illustrates the roles & responsibilities of different actors. Consider the limitations of evaluation methods laid out in the section on evaluation methodologies (p.32) and our case study (p.39) on evaluating misinformation that puts our evaluation framework into practice.
- **Public policy makers:** Look at figure 2.1 (p.10) that illustrates our three-layered evaluation framework, skim section 3 (p.12), which lays out the state of evaluation today; and read the part on roles and responsibilities (p.22) in the discussion section.
- **AI researchers:** Consider the evaluation framework in section 2 (p.7), concrete ways forward as introduced in section 4 and in the case study (p.39), and the limitations and implications in the discussion section (p.22).

1. 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 (Nori et al., 2023; Singhal et al., 2023) to news and politics (e.g. Bruell (2023)) and social interaction such as companionship (e.g. Griffith (2023); Pentina et al. (2023)). Early systems produced output in single modalities, such as image generation (Ramesh et al., 2021; Rombach et al., 2022) and text capabilities, producing compelling natural language (Anil et al., 2023; Glaese et al., 2022; OpenAI, 2023b). Increasingly, generative AI systems in other modalities such as audio, including voice and music (Agostinelli et al., 2023; Borsos et al., 2023; Dhariwal et al., 2020; Huang et al., 2023; Oord et al., 2016), and video and audiovisual capabilities are steadily improving (Du et al., 2023). Generative AI systems are increasingly multimodal, and their integration into various aspects of life is anticipated (Google Research, 2023).

In addition to creating benefits, generative AI systems pose risks of harm. For individual modalities, these risks have been mapped out in different taxonomies (Barnett, 2023; Bird et al., 2023; Bommasani et al., 2022; Dinan et al., 2021; Liu et al., 2023b; Shelby et al., 2023; Shevlane et al., 2023; Solaiman et al., 2023; Weidinger et al., 2021) as well as in research on individual risks or applications (e.g. Bianchi et al. (2023); Birhane et al. (2021); Carlini et al. (2023a); Khlaaf et al. (2022); Luccioni et al. (2023); Shevlane et al. (2023)). Complementing foresight research, observed instances of harm from generative AI systems have been logged to identify risks that these systems create (AI Incident Database; Organisation for Economic Co-operation and Development, b). Now that risks from generative AI systems have been identified, their impact on the overall safety of a generative AI system must be understood. This requires evaluation.

The growing use of generative AI systems makes it both easier and more pressing to evaluate potential risks of harm. As these technologies become widely used and embedded, the risks they create are a public safety concern. Accordingly, evaluating potential risks from generative AI systems is a growing priority for AI developers (Anthropic, 2023; OpenAI, 2023c), public policy makers (The White House, 2023), regulators (EU AI Act, 2023; National Institute of Standards and Technology, 2021a,b; UK Task Force), and civil society (Electronic Privacy Information Center).

Evaluation is the practice of measuring AI system performance or impact. Safety evaluation in particular focuses on evaluating risks of harm or actualised impacts on people or broader systems. Evaluations can be exploratory (such as open-ended probing of an AI system) or directed (such as running a specific test). They include qualitative investigations, such as studying how people actually attempt to use an AI system, as opposed to assessing intended use cases. Exploratory evaluations may identify areas of uncertainty or additional context, or give rise to novel directed evaluation questions. Directed evaluations follow a series of steps, whereby a target – such as a risk of harm – is selected, operationalised into an observable metric, and measured. In any evaluation, the results are then judged against a normative baseline, such as whether an AI system is "good", "fair", or "safe enough". Evaluation is never neutral: it rests on interwoven technical and normative decisions, such as deciding what to evaluate in the first place, how to measure it (see Operationalising

[1] By "generative" we refer to AI systems that generate novel output rather than analysing existing data (Huang et al., 2022). We focus on generative AI systems, which we define as models that input and output any combination of image, audio, video, and text. This includes transformer-based systems, such as large language models, diffusion-based systems, and hybrid architectures.

[2] By "multimodal" we refer to models that accept and produce output in any combination of image, audio, and text. This includes models that accept or produce output in more than one modality, such as interleaved image and text data, or audiovisual data.

[3] By "AI system" we refer to a pre-trained base model or foundation model, potentially "fine-tuned" by adapting it to particular datasets for specific performance targets, including via practices such as RLHF. AI systems may also include filters such as input or output filters. AI systems are ready for integration into a product.

risks), and what results indicate "good" AI system performance (Bakalar et al., 2021; Bowker and Star, 2000). Safety evaluation can form part of broader safety audits, which may additionally take into account organisational governance structures or existing documentation and more (Costanza-Chock et al., 2022; Mökander et al., 2023; Raji et al., 2020).

Evaluation performs an important function by providing public safety assurances. By systematically testing AI systems against potential risks of harm, evaluation can make AI systems less opaque. Evaluation also sheds light on, predicts, and quantifies the likelihood of potential downstream harms, and can surface the factors and mechanisms that influence whether downstream harm may occur. Evaluations can guide the development of AI systems, as well as providing assurances on levels of AI system safety in different contexts. As a result, the understanding of AI systems that evaluations provide is essential for well-informed, responsible decision-making on AI system development and deployment (Stilgoe et al., 2013). Further, evaluation of different risk areas brings to light normative trade-offs that arise as AI systems are developed and deployed in real-world settings. By performing these functions, evaluation is a foundation for meaningful accountability on the responsible innovation and deployment of generative AI systems.

In this paper, we make two main contributions: a sociotechnical framework for safety evaluation and an empirical assessment of the current safety evaluation landscape. While the priority of safety evaluations for generative AI systems is clear, current approaches are often heterogeneous and ad hoc. The evaluations that are being conducted differ between organisations and AI systems (e.g. Anil et al. (2023); Anthropic (2023); Glaese et al. (2022); Mishkin et al. (2022); OpenAI (2023a)), which makes them hard to compare and reproduce. This can also mean that the evaluation of a given AI system misses important risks that should be considered. We argue that a more systematic and standardised approach to safety evaluation is necessary to ensure meaningful, comparable, and comprehensive safety evaluation (c.f. Liang et al. (2022); National Institute of Standards and Technology (2019); Vogel and Manyika (2023)). As a step in that direction, we offer a sociotechnical framework to guide safety evaluation of generative AI systems.

Our second main contribution is a review of the current state of safety evaluations for generative AI systems, including the public release of a repository of existing evaluations, and analysing gaps. These gaps are tractable: we present practical steps toward closing them. We propose roles and responsibilities for different stakeholders and discuss how currently disparate communities with interests in the safety of AI systems can intersect.

The paper proceeds as follows. In section 2, we outline our proposed framework for sociotechnical safety evaluation across three layers that progressively add context: the capability layer, the human interaction layer, and the systemic impact layer. Section 3 surveys the current state of research and practice in sociotechnical safety evaluation, identifying strengths and limitations of existing approaches. Section 4 builds on the framework and survey to discuss ways to close observed evaluation gaps. In section 5, we conclude with a discussion of open questions for the field of safety evaluations, including the varying roles and responsibilities between AI developers and public sector stakeholders for conducting evaluations across the layers, and the connections between the range of proposed approaches to safety associated with generative AI systems.

2. Framework for sociotechnical AI safety evaluation

Recent research identified a sociotechnical gap in our understanding of the safe development and deployment of AI systems (Lazar and Nelson, 2023; Mohamed et al., 2020; Selbst et al., 2019; Shelby et al., 2023). This sociotechnical gap arises where AI system safety is evaluated only with regard to technical components of an AI system, i.e. individual technical artefacts such as data, model architecture, and sampling

strategies. While these are important aspects of AI safety evaluation, they alone are insufficient to determine whether an AI system is safe. Instead, an approach is needed that takes into account human and systemic factors that co-determine risks of harm.

To close this gap, we apply a sociotechnical lens to AI safety evaluation. Sociotechnical research has a long-standing history in expanding the frontiers of AI system evaluation to include human and systemic factors (Barocas and Selbst, 2016; Dwork et al., 2012; Ekstrand et al., 2018; Friedman and Nissenbaum, 1996; Raji et al., 2020). This approach is rooted in the observation that AI systems are sociotechnical systems: both humans and machines are necessary in order to make the technology work as intended (Selbst et al., 2019). The interaction of technical and social components determines whether risk manifests (Leveson, 2012). Consequently, AI evaluation requires a framework that integrates these components and their interactions.

Similarly to other sociotechnical work (c.f. Raji et al. (2020); Rismani et al. (2023)), our approach is further inspired by a system safety approach from the discipline of safety engineering. System safety represents a historical paradigm shift in safety engineering, from component-based approaches toward systems thinking, taking into account broader contexts, interactions and emergent properties of complex systems (Leveson, 2012). Component-based approaches to safety emerged historically from industries dealing with hazardous materials in constrained settings and processes. These approaches isolate system components or steps in a process and assess individual failure modes of each part. The sum of these assessments is then considered a comprehensive safety evaluation of the entire system. This approach is not fit for purpose in complex and versatile systems, such as where software and people interact with great degrees of freedom (Hutchins, 1995; Leveson, 2012). Here, a component-based approach to safety needs to give way to a system-based approach to safety (Leveson, 2012). An analogous shift is required in the safety evaluation of generative AI systems. We propose a framework that accounts for this shift.

Specifically, we present a three-layered framework to structure safety evaluations of AI systems. The three layers are distinguished by the target of analysis. The layers are: *capability* evaluation, *human interaction* evaluation, and *systemic impact* evaluation. These three layers progressively add on further context that is critical for assessing whether a capability relates to an actual risk of harm or not.

To illustrate these three evaluation layers, consider the example of misinformation harms. *Capability* evaluation can indicate whether an AI system is likely to produce factually incorrect output (e.g. Ji et al. (2023); Lin et al. (2022)). However, the risk of people being deceived or misled by that output may depend on factors such as the context in which an AI system is used, who uses it, and features of an application (e.g. whether synthetic content is effectively signposted). This requires evaluating *human-AI interaction*. Misinformation risks raise concerns about large-scale effects, for example on public knowledge ecosystems or trust in shared media. Whether such effects manifest depends on *systemic* factors, such as expectations and norms in public information sharing and the existence of institutions that provide authoritative information or fact-checking. Evaluating misinformation risks from generative AI systems requires evaluation at each of these three layers.

The three layers in this framework interact and their boundaries are gradual. Effects detected at one layer may be indicate related observations at the next. For example, discrepancies in how an AI system performs for different user groups can be identified at the layer of human interaction, and may foreshadow disparate systemic impacts for these groups. A further illustration of the gradual boundaries between the layers is that evaluation methods can straddle multiple layers. For example, adversarial testing is a method for evaluating capabilities. However, by focusing on the experience of the adversarial tester, it can be a measure of human interaction: specifically of the friction a person encounters when trying to use an AI system to malicious ends. While interactions between these layers may extend beyond the

failure of individual system components and be complex, they are often still within the control of AI system developers.

In addition, there are feedback loops within and between layers. For example, societal context may feed back into system capabilities via the opinions and demographics of human annotators, as annotated data is used to adapt AI systems to particular contexts. Note that the layers are not ordered by importance nor in any chronological order. Rather, evaluations at each layer can be performed simultaneously and asynchronously. We now introduce the three-layered framework in detail.

2.1. Layer 1: Capability

Evaluation at this layer targets AI systems and their technical components.[4] These are routinely evaluated in isolation, including tests of AI system behaviour in response to novel tasks or environments, or testing individual technical artefacts, such as the data that an AI system is trained on. It also includes evaluating processes by which these artefacts are created, such as the aggregation mechanisms in processes that are used to adapt an AI system to a particular task. In addition to assuring the safety of an AI system, evaluations at this layer are often performed to guide iterative model development ("hill-climbing").

While evaluation at this layer does not assess downstream harm per se, it can provide indication of whether a component, output, or AI system is likely to cause downstream harm. Several risks of harm can be evaluated by measuring capabilities through the outputs of an AI system. This includes, for example, the extent to which an AI model reproduces harmful stereotypes in images or utterances (representation harms (Bianchi et al., 2023)), makes factual errors, or displays advanced capabilities that present safety hazards. The extent to which model performance deteriorates when prompted in different languages, about different groups, or in different domains can also be evaluated at this layer and can be indicative of the likely distribution of potential downstream harm. Capabilities also include metrics that are designed to track efficiency and may shed light on potential downstream environmental impact, such as energy use at inference (Kaack et al., 2022). Capabilities can be assessed against fixed, automated tests or probed dynamically by human or automated adversarial testers (see Selecting evaluation methods).

Evaluations at this layer can also concern the data on which a model is trained. Using tools to visualise clusters or associations in the training data can reveal diversity and representativeness of the data, or the presence of sensitive data such as private information (Choi et al., 2023; Dodge et al., 2021; Kreutzer et al., 2022; Wang et al., 2022). Similar tools can be used to assess the learned associations of a trained AI system (Caliskan et al., 2017; Steed and Caliskan, 2021).

Other components that can be analysed at this layer include filters and techniques used to reduce output that may relate to a particular risk of harm, such as filters for safety harms in images (Rando et al., 2022) or toxic language (Perspective API). However, such filters have limitations (Rando et al., 2022) that can aggravate representation harm by disproportionately filtering out content from some groups (Welbl et al., 2021).

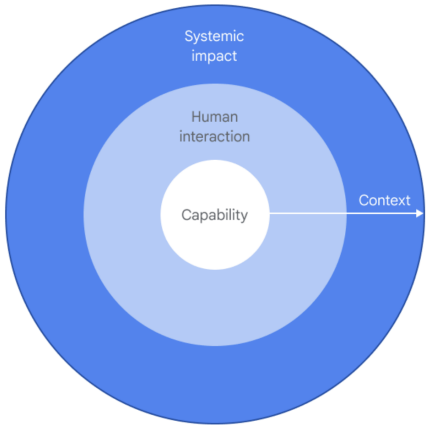
Capability evaluation is critical, but insufficient, for a comprehensive safety evaluation. It can serve as an early indicator of potential downstream harms, but to assess whether or not a capability relates to risks of harm requires taking into account context – such as who uses the AI system, to what end, and under which circumstances. This context is assessed at subsequent layers.

2.2. Layer 2: Human interaction

This layer centres the experience of people interacting with a given AI system. Assessing AI system safety requires evaluating not just the AI system in isolation but also effects on people interacting with AI systems, and the human–AI

[4] These include training data; model components, such as model architectures and classifiers; the model itself, such as pre-training embeddings; and model outputs, such as images.

Figure 2.1 | A sociotechnical framework for safety evaluation comprises three layers.



dyad. This includes usability: does the AI system perform its intended function at the point of use, and how do experiences differ between user groups? This layer also centres potential externalities: does human–AI interaction lead to unintended effects on the person interacting or exposed to AI outputs? Evaluation at this layer acknowledges that AI system safety depends on who uses an AI system, with what goal in mind, and in what context. This layer shifts the lens to the humans interacting with an AI system and is key to a human-centred approach to AI development (Liao and Vaughan, 2023; Tahaei et al., 2023; Vaughan and Wallach, 2021).

In addition to testing AI system capabilities, the functionality of an AI system in the context of a concrete application must be assessed (Raji et al., 2022a). This includes testing how different people actually use the system, as real-world use often deviates from intended use cases. User groups are heterogeneous, and safety evaluation requires not only assessing whether an AI system works but also for whom it works well (Wang et al., 2023). To assess usability in practice, human interaction with AI systems needs to be evaluated "in the wild", i.e. in a real-world application context for an AI system such as a hospital (Sendak et al., 2020) or police units (Marda and Narayan, 2020). Evaluating the human–AI interaction can also reveal how easy it is to use a model for malicious ends (Roy and Umbach, 2023).

Human-centred testing can shed light on potential externalities created by specific use cases or applications of AI systems. To assess a risk of harm, directed psychology or human–computer interaction experiments can be performed. Under controlled, safe conditions, potential harmful outcomes to people interacting with AI can be studied, such as overreliance on AI systems (Chiesurin et al., 2023) or overtrust (e.g. due to AI systems endowed with anthropomorphic cues, Glikson and Woolley (2020)). Some effects may only manifest over time and require longitudinal evaluation. For example, one experiment found that repeated exposure to AI content increases its persuasiveness but only up until a certain point (Cacioppo and Petty, 1980). Regarding another risk area, it has also been hypothesised that increased feelings of social isolation due to overuse of technology may only show up after frequent exposure to an AI system, in between a user's interactions (Turkle, 2011). Evaluation at this layer can also assess harms to data annotators, as they are exposed to harmful model outputs, including via surveys or interviews (Gray and Suri, 2019; Stoev et al., 2023). In addition, human interaction evaluations may reveal disparate harm profiles for different modalities. For example, one set of evaluations found that users more readily believe synthetic misinformation that is presented in video as opposed to text modalities (Sundar et al., 2021).

Evaluations at this layer can also identify psychological mechanisms by which harms may occur to a person interacting with an AI system. For example, they may identify cognitive biases that influence people coming to believe misinformation (Jerit and Zhao, 2020), or how AI systems influence or persuade humans over the course of an interaction, such as when co-writing a text (Hohenstein et al., 2023; Jakesch et al., 2023).

Finally, evaluation that considers an AI system in the context of use can assess the overall performance of the human–AI dyad, such as quality of outcomes on AI-assisted computer coding tasks compared to a human–human

baseline (Vasconcelos et al., 2023).

While this layer provides critical context by adding human interaction to the evaluation, it remains insufficient for a comprehensive AI safety assessment. It provides limited insights on the potential broader impacts that an AI system may have when deployed at scale, and does not consider risks and impacts on broader systems such as society, economic impacts, or the natural environment. Assessing these effects requires analysing the broader systems into which an AI system is deployed, at the third and final layer of our sociotechnical framework for safety evaluation.

2.3. Layer 3: Systemic impact

The third target of evaluation is the impact of an AI system on the broader systems in which it is embedded, such as society, the economy, and the natural environment. Widely used AI systems shape, and are shaped by, the societies in which they are used (Matias, 2023; Wagner et al., 2021). Detecting the effects from these interactions requires evaluation at the system layer. Some effects may only emerge as an AI system is deployed at large scale. For example, risks from increasing homogeneity in knowledge production and creativity due to "algorithmic monocultures" are emergent at the systems layer of evaluation (Doshi and Hauser, 2023; Kleinberg and Raghavan, 2021; Toups et al., 2023). Harms may also have small effect sizes that are hard to detect at the layer of individual user interactions but become salient at a systems level (e.g. Bulimia Project). Evaluating these risks and impacts requires focusing on the broader systems into which the AI system is integrated.

Evaluation at this layer can target systems of different domains and sizes. Economic assessments may concern broad systemic impacts, such as the labour market impacts of generative AI (Eloundou et al., 2023; Felten et al., 2021; Frank et al., 2019; Frey and Osborne, 2013; Tolan et al., 2021) or the impact of model adoption on productivity (Brynjolfsson et al., 2023). It may also centre specific industries or goods, for example by evaluating impacts on the creative economy or predicting the likely impacts of generative AI on the erosion of public goods such as the creative commons (del Rio-Chanona et al., 2023; Huang and Siddarth, 2023). Impact from generative AI systems on societal institutions, such as political polarisation or changes to trust in public media, can be evaluated through system evaluation (Lorenz-Spreen et al., 2023). The fairness of how benefits and risks are distributed can also be ascertained at this layer, for example by assessing take-up of AI tools across countries (Calvino and Fontanelli, 2023) and identifying who is able to capture and extract most value using these technologies (Brynjolfsson et al., 2023).

Evaluations at this layer may also focus on smaller, more localised systems, such as assessing impacts from an AI system in a clinical context on the provision of care (Elish and Watkins, 2020). Evaluating how AI systems are socially embedded can shed light on how people come to trust the outputs – for example, where friends and colleagues all use a system, this system may be trusted more. One common concern with the widespread availability of generative AI systems is that they can be used to cheat on school assignments (Rudolph et al., 2023). System-level evaluation of adoption and perception of AI systems can evaluate what types of use occur and under what circumstances, and by whom they constitute 'cheating'. Environmental impacts can be targeted at this layer, to provide a nuanced understanding of impacts on broader ecosystems. For example, detailed and localised evaluation can reveal the actual environmental impact from generative AI systems, such as from data centres that rely on nearby water sources for cooling (Luccioni et al., 2022). Early stage indicators, such as energy use as a proxy for environmental impact at the capability layer, can be calibrated and contextualised via evaluations at this layer.

Evaluation at this layer can also provide context-rich assessments of interactions of different systems as social, economic, and ecological factors intersect. For example, evaluation at the system layer may take into account the biodiversity and resilience of local ecosystems, the nature of the energy grid, and

the social and economic implications for local communities in order to assess overall harm of infrastructure that powers an AI system (Solaiman et al., 2023).

Systemic impacts are often difficult to assess due to the complex nature, idiosyncrasies, and noise of the systems that are being evaluated. While direct impacts of an AI system may not be known until post deployment, forecasts or comparable technologies can provide initial insights on potential risks of harm at this layer.

2.4. Summary

We present a three-layered sociotechnical framework for safety evaluation of generative AI systems. (While we focus on generative AI systems, this framework may also be applicable to other types of AI.) The same high-level risk areas can be detected and evaluated at each layer (we outline practical steps toward evaluating risks at each layer in section 4). What connects the three layers is that they progressively add on further context. Note that they are not sequential or dependent on each other. Neither are the layers conditional on each other; rather, evaluation at each layer can be run simultaneously. Integrating results from all layers provides a comprehensive evaluation of the safety of a generative AI system. The layers are a guiding structure to facilitate evaluation along different layers of context in a sociotechnical system.

3. Current state of sociotechnical safety evaluation

In this section, we survey the state of safety evaluation of generative AI systems. This first requires consolidating a taxonomy of potential harm from these AI systems. We present a synthesised taxonomy of harm based on prior literature of taxonomies of harm from generative AI systems. Next, we employ an extensive process to identify all existing evaluations of generative AI systems that speak to risks identified in our taxonomy. We map all identified evaluations by risk area; by AI system modalities (image, audio, video, text, and multimodal combinations); and by layers of evaluation based on the three-layered framework introduced above. This mapping is presented in an overview figure that snapshots the sociotechnical evaluation landscape today. We close this section by discussing the "evaluation gaps" this mapping has surfaced.

3.1. Taxonomy of harm

First, to assess the state of sociotechnical safety evaluation for generative AI systems requires grounding the types of risk that such evaluation should assess. To this end, we revisit the growing literature on social, ethical, and other safety risks from generative AI systems and integrate insights from this literature into a single, holistic taxonomy (high-level version, table 1; detailed version, appendix section A.1). Rather than presenting a novel research artefact, the goal of this taxonomy is to provide a basis for mapping the state of safety evaluation of generative AI systems.

Previous work identified a wide range of risks posed by generative AI systems. Existing taxonomies address risks from AI systems audio (Barnett, 2023) and text (Bommasani et al., 2022; Liu et al., 2023b; Weidinger et al., 2021), as well as combined modalities, such as text-to-image (Bird et al., 2023). Solaiman et al. (2023) provide an overview of harms from generative AI systems writ large and describe approaches to social impact analyses for each identified harm area. In our overview, we include both established and emerging risks. Established risks are defined by observed instances of harm, such as representation risks (e.g. Bianchi et al. (2023); Birhane et al. (2021); Luccioni et al. (2023)). Emerging risks are anticipated based on the foreseeable capabilities of generative AI systems, such as increasingly persuasive content produced by generative AI (Matz et al., 2023; Shevlane et al., 2023).

We build on this prior literature to aggregate a single, holistic taxonomy of harm from generative AI systems. This taxonomy has six high-level harm areas: 1. Representation & Toxicity Harms, 2. Misinformation Harms, 3. Information & Safety Harms, 4. Malicious Use, 5. Human

Autonomy & Integrity Harms, 6. Socioeconomic & Environmental Harms (see table 1).

We present a high-level overview of this taxonomy below (table 1), with examples of how these risks may manifest in modalities other than text. A more detailed breakdown of each risk area is provided in appendix section A.1.

3.1.1. Multimodality raises new evaluation challenges

While none of the higher-level risk areas are new in multimodal as opposed to text-based generative AI systems, the specific ways in which they may manifest are likely to differ between modalities. For example, violent or sexually explicit content has a greater "shock factor" in image modalities than in text. Multimodal models may also introduce novel evaluation challenges. For example, consider a text-to-image AI system that produces images based on text input. Hutchinson et al. (2022) argue that text often underspecifies context such that an image of a "wedding", for example, will necessarily include certain objects and cultural contexts, regardless of whether these concepts are articulated in the accompanying text. It may be easier for an AI system to hedge or give pluralist output in text than in images.

Risks may also 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., 2021b). Similarly, an innocuous video of military training exercises combined with audio describing the invasion of a country risks creating an instance of misinformation (Vincent, 2023). Generative AI systems may also perpetuate stereotypes in ways that are highly dependent on domains – for example, by overrepresenting nude females as compared to nude males in the context of synthetic music videos. Detecting these harms may build on existing methods but is likely to require novel context-sensitive evaluation approaches. Though risk evaluation will likely draw from lessons learned evaluating models in single modalities, novel evaluations that enable a holistic view across modalities are required to capture risks in multimodal AI systems.

3.2. Mapping the landscape

We now present the results of a large-scale review of existing benchmarks to assess risks of harm from generative AI systems. To write this section, our group of co-authors and reviewers assembled an overview of all sociotechnical safety benchmarks and evaluation methods known to this group up to 10 October 2023.

Included are academic papers or online reports that meet two criteria: they constitute an evaluation, and they have been applied to a generative AI system. 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 describes results from its application to a generative AI system. Note that evaluations that may be applicable to generative AI systems but have not been applied to such systems yet were not in scope for this review.

All submitted evaluations were coded based on output modality – whether they evaluate text, image, audio, video, or multiple modalities. They were then coded based on what risk area they cover, based on the above taxonomy. Finally, they were coded by layers of evaluation in our three-layered framework (1-capability, 2-human interaction, 3-systemic impact).

We release an evaluation repository of all included evaluations as an open resource here. The contents of this survey evaluation repository are presented in an overview figure (figure 3.1).

3.2.1. Limitations

While great efforts were made to conduct a large-scale review of existing evaluation approaches, we do not assume that this mapping is comprehensive. Our approach is further limited by not considering input

Table 1 | High-level overview of risks of harm from generative AI systems

Harm 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., 2023a)
Misinformation Harms	AI systems generating and facilitating the spread of inaccurate or misleading information that causes people to develop false beliefs	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 leaking, reproducing, generating or inferring sensitive, private, or hazardous information	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 compromising human agency, or circumventing meaningful human control	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 employment, innovation, and the environment	Exploitative practices to perform data annotation at scale where annotators are not fairly compensated (Stoev et al., 2023)

modality: our coding is based on output modality. Future mappings may distinguish between input modalities for a more fine-grained analysis (e.g. mapping 'image-to-text' evaluations distinctly from 'text-to-text' evaluations). Finally, our mapping is a snapshot of a moment in time. In the future, it may be conducive to a thriving ecosystem on sociotechnical evaluations to expand the evaluation repository into a living resource that evaluation developers can add their methods to.

3.3. Evaluation gaps

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

1. **Coverage gap: Evaluations for several risks are lacking.** Coverage of ethical and social risk evaluation overall is low. Several gaps exist where there are few or no evaluations to assess a given risk area.
2. **Context gap: Human interaction and systemic evaluations are rare.** Existing evaluations cluster in the text modality, with fewer evaluations available for audio, image, video, or combinations of modalities. This presents a challenge for evaluating social and ethical risks in other modalities.
3. **Multimodal gap: Evaluations are missing for multimodal AI systems.** Most evaluations of social and ethical harms that were identified cluster at the layer of capability evaluations. We now discuss these observations in turn.

First, we observe that evaluations are scarce for several previously identified risks from generative AI systems. This lack of coverage is particularly pronounced for information and safety harms, human autonomy and integrity harms, and socioeconomic and environmental harms. While the number of available evaluations is insufficient to assess coverage of a risk area, the absence of evaluations is a clear signal that the given risk area cannot currently be evaluated in generative AI systems.

More detailed inspection of the evaluation repository indicates that the lack of coverage extends beyond these three risk areas: even where evaluations exist, they do not cover the risk area comprehensively. For example, we identified 83 evaluations of representation harms. However, these cover only a small space of representation harms – 17% of them cover binary gender and

Figure 3.1 | Evaluations per harm area and AI system output modality. No harm area is well covered across modalities.

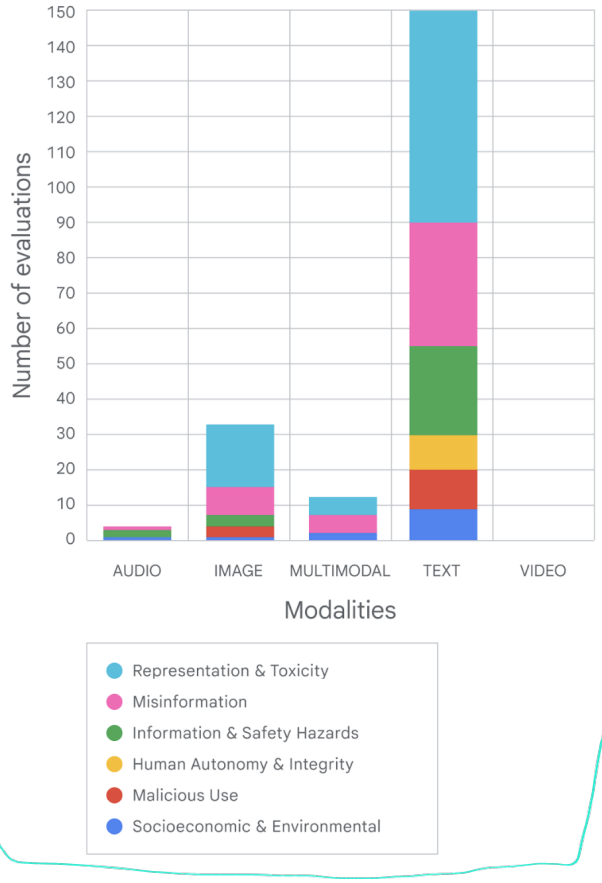
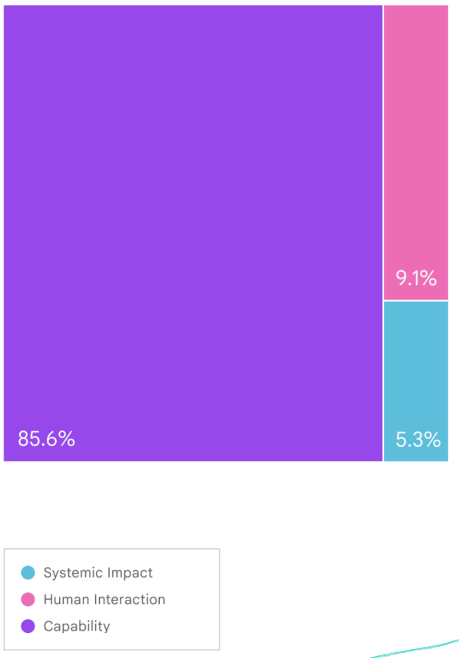


Figure 3.2 | Evaluations per layer. Human interaction and systemic impact evaluations to assess generative AI system safety are rare.



occupation bias,[5] and 60 cover text modalities only. They also cover only a small space of the potential harm: multiple "discriminatory bias" benchmarks cover binary gender or skin colour as potential traits for discrimination (Cho et al., 2023; Mandal et al., 2023). These evaluations do not cover potential manifestations of representation harms along other axes such as ability status, age, religion, nationality, or social class. In sum, further evaluations are needed to cover ethical and social harms, including plugging more nuanced gaps in risk areas for which some evaluations exist.

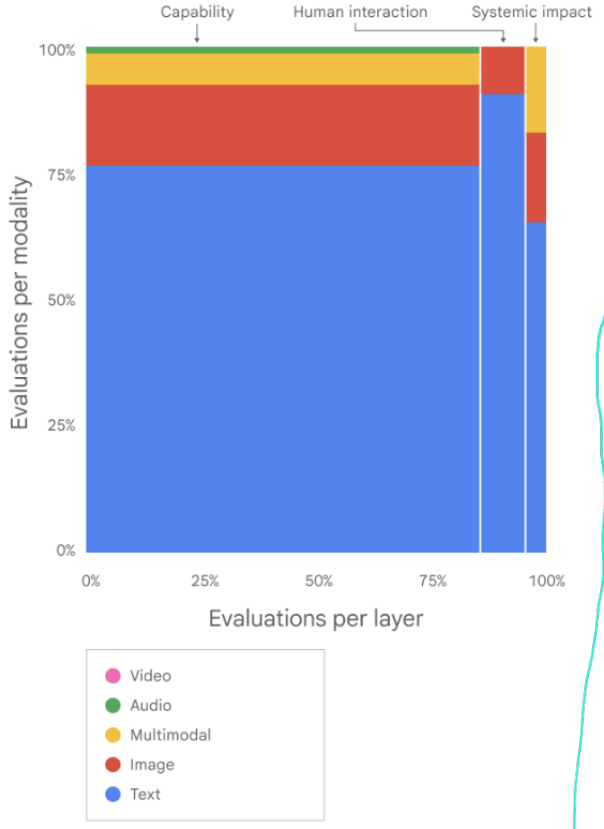
Our second main observation is that insofar as evaluation tools exist to address risks from multimodal generative AI, they are mainly clustered at the capability layer. More detailed inspection of the repository indicates that evaluations focus particularly on AI system outputs and to a lesser degree on available training data. This clustering of evaluation at the capability layer is reflective of, and likely partially driven by, the evaluations that have recently been performed and disclosed as part of large generative AI system announcements, which primarily focus on capability evaluations (Anil et al., 2023; Anthropic, 2023; Glaese et al., 2022; Mishkin et al., 2022; OpenAI, 2023a; Touvron et al., 2023).

While a capability-focused approach provides important indications as to potential downstream harms, it does not account for contextual factors that co-determine risks of harm (see section 2). Capability evaluation is a core piece of safety evaluation, but it must be complemented by further analyses that add layers of relevant context. As a result, further work is needed to expand sociotechnical evaluations at the human interaction layer and at the system layer.

Our third observation is that the vast majority of evaluations exclusively assess text. Few evaluations exist for image outputs or

[5] Six of the fourteen focused exclusively on gender and occupation. The rest include additional demographics and stereotypes.

Figure 3.3 | Evaluations per layer and modality. Most (75%) of all evaluations target text output.



combinations of text and image, and evaluations of audio or video modalities are scarce. There are only four publicly documented evaluations targeting audio and we did not find any evaluations targeting video.[6] This may in part be a result of historical contingencies: generative AI systems that output text saw rapid, widespread adoption, which may have triggered proportionately more research into ethical and social risks and corresponding evaluations.

Generative AI systems that produce compelling audio including voice and music already exist (Agostinelli et al., 2023; Borsos et al., 2023; Dhariwal et al., 2020; Huang et al., 2023; Oord et al., 2016), and 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 audiovisual video with subtitles – creates different manifestations of harm across the six identified harm areas. As a result, assessing ethical and social harm in multimodal models requires novel evaluation approaches. (We discuss some steps toward this in section 4.)

Critically, this distribution of evaluations centring text modalities is not driven by a principled assessment of the modalities in which harm is likely to occur. Several risks have been anticipated in the audio, image, and video modalities or combinations (see appendix section A.1). 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 no evaluation tools exist to assess this in generative AI systems. In some cases, evaluations designed for text output can be repurposed for other modalities (see section 4). However, this is limited, especially where the same risk may manifest differently across modalities.

Combinations of modalities can create novel risks as well as compound effects. For example, misinformation has been found to be more compelling in audiovisual modalities as opposed to text (Hameleers et al., 2020). AI systems that span multiple modalities may also be more vulnerable to malicious attacks aimed at getting a model to create harmful output, as fewer safety mechanisms and less exploration of vulnerabilities exist for them (Carlini et al., 2023b). Thus, evaluations must be expanded to modalities other than text. In addition to evaluating individual modalities in isolation, they must also be expanded to assess compositions of modalities, i.e. multimodal outputs.

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

4. Closing evaluation gaps

Our assessment of the current state of safety evaluations of generative AI systems identified significant gaps. In this section we propose practical steps to close these gaps. These steps are tractable. Closing these gaps will require work. In addition, it may involve clarifying roles and responsibilities, which we return to in section 5. This section is primarily aimed at practitioners and those funding or performing the construction of new evaluations.

To close identified gaps, new evaluations are needed. In part, this likely means constructing novel evaluations. The first part of this section presents the general pipeline and building blocks for constructing such evaluations. In particular, we describe how rich, multifaceted concepts of harm can be made measurable through the process of "operationalisation" (see Operationalising risks). We then outline concrete methodologies that can be used to obtain measures of a given AI system, for each layer of evaluation (see Selecting evaluation methods).

In addition to constructing novel evaluations, it may be possible to extend existing evaluations to generative AI systems. The second part of this section focuses on practical steps that can be taken to close the gaps in the evaluation of generative AI systems. We discuss these practical avenues and their advantages and limitations in the second part of this section (see Practical steps to closing gaps in safety evaluation).

4.1. Operationalising risks

Evaluation is a process involving several steps: it requires first selecting a target (such as a risk of harm, e.g. "bias"); then operationalising it into a concrete metric (e.g. the association of gender and occupation, Luccioni et al. (2023)); then obtaining a measurement; and finally judging the outcome. Each of these steps has technical and normative elements (see Evaluation is never value-neutral). In this paper, we lean on previous literature to identify target constructs – namely, a taxonomy of identified risks of harm (see appendix section A.1). But how to proceed from a complex, multifaceted concept such as "misinformation" to a valid, tractable measurement of this risk? The process of operationally defining risks of harm such that they can be measured is the focus of this section.

Risks of harm from generative AI systems are often latent constructs that are not directly observable via a single test or metric (Jacobs and Wallach, 2021).[7] In order to measure these risks, they need to be operationally defined (King et al., 2021). Operationalisation is the process of mapping tractable, observable metrics or concepts to latent constructs. Measurement on these observable metrics is then taken to provide insight on the latent target construct. Note that operationalisation is inherently an ambiguous process. What constitutes a valid measure of harm is a contestable decision, and often metrics are iterated on and improved over time (Chang, 2004). Operationalisation may also constitute normative trade-offs – for example, on how a single performance metric should weigh false positives against false negatives. Operationalisation of complex constructs creates various pitfalls that can result in invalid measurements (see Operationalising risks).

In section 2 above, we argued that risks of harm from generative AI systems cannot be comprehensively assessed at a single layer of evaluation. Rather, complementary evaluation at all three layers is needed for a full evaluation. Thus, we propose operationally defining harm constructs at each of the three layers of evaluation.

Specifically, this requires mapping metrics or concepts that are observable at a given layer to the latent harm construct. Different aspects of a risk can be measured at each layer. Correspondingly, different metrics can be mapped to a given risk per layer. Metrics can range from single, observable, narrow metrics (e.g. the FID score to assess the quality of a generated image) to more open-ended empirical or qualitative metrics (e.g. user preferences or broader societal impact).

[7] This also applies to other targets of evaluation, such as cognitive capacities or potential benefits of AI systems.

For example, to operationalise the risk of information and safety harms, we may define the following metrics at each layer. At the capability layer, we assess properties of model output that indicate potential information hazards, such as the capability to output harmful biological information (Anthropic, 2023; OpenAI, 2023a). At the human interaction layer, this risk can be operationalised in multiple ways, one of which might be the likelihood of people unintentionally following instructions to assemble dangerous compounds in different contexts, e.g. where the generative AI system is used as a laboratory assistant. At this layer, the risk likelihood may also be measured via the friction that people encounter when intentionally trying to apply dangerous AI capabilities to malicious ends. Finally, at the systemic impact layer, this risk can be assessed via modelling potential distribution mechanisms of novel biohazards created based on such information. (These are examples for illustration purposes; for a more in-depth example, see the operationalisation of misinformation harms in our Case study: Misinformation.)

4.1.1. Ensuring validity

Some information inevitably gets lost when operationalising complex constructs 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. There are different ways in which validity may be compromised. We briefly canvass these and outline approaches to mitigating validity concerns.

Tests often do not measure precisely what they seek out to measure: they may capture only a subset or part of the target construct (internal validity), or may capture the phenomenon fully in a given instance but not allow extrapolation to new situations (external validity) (Liao et al., 2021).

AI capability evaluation in particular has been criticised for relying on overly narrow operational definitions of complex harms, leading to both internal and external validity failures (Liao et al., 2021; Raji et al., 2021). Operational definitions must be arrived at carefully and deliberately, or they risk yielding misleading results. For example, one study found that the risk of harmful stereotyping in language modelling had been operationalised as the association of word pairs, but only some of the referenced word pairs were actually harmful and others were innocuous, such as the word pair "Norwegian" and "salmon". As this operationalisation included instances that were not harmful, the validity of the resulting metric and what it can say about harmful stereotyping was fundamentally called into question (Blodgett et al., 2021). Similar validity failures have been exposed in other evaluation approaches, particularly in narrow tests such as automated benchmarks (Rauh et al., 2022; Schlangen, 2019), which we return to later in this section (see Selecting evaluation methods).

To mitigate such validity issues, multiple approaches can be taken. Specific validity challenges for individual methods are described in detail in Selecting evaluation methods below. Here, we outline general best practices to assure the validity of a given evaluation:

- **Grounding the operationalisation of a risk of harm.** An evaluation can, for example, be grounded in a literature review of a given harm, or in human annotation or examples curated by experts. To stress-test definitions and operationalisations, invite diverse perspectives and multiple lenses onto the same risk of harm. Participatory, expert-led, and interdisciplinary approaches can be helpful here (e.g. Narayanan and Kapoor (2023)).
- **Documenting and signposting limitations of a given evaluation.** As risks of harm are latent concepts, no single operationalisation captures them in their entirety. By making choices on how to operationalise a given risk explicit and documenting them, others can better interpret results and identify limitations (Mattson et al., 2023; Raji et al., 2021).
- **Cross-validating an operationalisation by comparing results from different**

evaluations of the same concept. If the results do not align, this indicates areas in which metrics operationalise a harm in divergent ways (e.g. Goldfarb-Tarrant et al. (2021)).

- **Making results interpretable.** This may include aggregating multiple results into a single, overall result that captures multiple facets of a harm. However, collapsing multiple tests into a single result requires care as it can also make it harder to identify validity failures of individual items.

4.2. Selecting evaluation methods

Once a risk of harm is operationally defined, appropriate methods must be selected to obtain these measurements. Often, the selection of evaluation methods is intimately entwined with defining the metrics. In this section, we describe available methods for measuring risks of harm from generative AI systems at each of the three layers of evaluation. For each method, we provide examples of sociotechnical evaluations, and discuss methodological limitations in appendix section A.2.

4.2.1. Capability evaluation methods

To assess model capabilities, practitioners may leverage *automated evaluations* that assess performance against fixed datasets or tasks. Alternatively, *human annotation* can evaluate AI system capabilities against specified goals or failures, such as whether an image includes violent images. Human data annotation can be used to develop novel automated evaluations. Capability testing can be *adversarial*, whereby humans or automated tools may probe a model to identify pathways that lead to failure modes. Such adversarial probing can be quite exploratory and may, in some instances, surface unexpected risks or failure modes.

Evaluation methods at this layer can be grouped as follows. We provide detailed descriptions, examples, and a discussion of limitations in appendix section A.2.1:

- Human annotation
- Benchmarking
- Adversarial testing

4.2.2. Human interaction evaluation methods

This layer centres the experience of humans interacting with AI systems. Evaluation at this layer always requires human participants, as their experiences and effects or externalities on human interactants are the subject of study.[8] The extent to which AI systems influence or shape human preferences and behaviours can be assessed via *behavioural experiments*. These experiments can bring general mechanisms and effects into focus. Assessing the consequences of specific features, use cases, or application domains requires user *research*. User studies can also assess how people actually attempt to use a generative AI system, as contrasted with the use case intended by designers. Whether an AI system functions across domains and how it performs for different user groups is core to a range of social and ethical risks, and can be assessed through user testing. While behavioural experiments and user testing require some abstraction from real-world use, passive monitoring of how people use deployed systems can provide insights on downstream effects in real-world contexts. Mixed-methods approaches that integrate different sources of data, such as behavioural observations and survey data, often provide the most robust results. Some AI system impacts on users or interaction effects may manifest only over the course of prolonged or frequent interaction; detecting these requires longitudinal designs of any of these research methods that evaluate human–AI interaction over time.

Evaluation methods at this layer can be grouped as follows. We provide detailed descriptions, examples, and a discussion of limitations in appendix section A.2.2:

- Behavioural experiments
- User research

[8] It has been proposed to simulate human participants in social science research (e.g. Argyle et al. (2023); Dillion et al. (2023)), but these methods are in their infancy – i.e. in the early, exploratory stages. They cannot therefore be relied upon for robust information to underpin responsible decision-making in AI system development.

- Passive monitoring of human use

4.2.3. *Systemic impact evaluation methods*

At the system layer, evaluation methods target the emergent effects from interactions within the sociotechnical system of which an AI system is part. This includes *staged release* or *pilot studies* and *ex-post impact assessments* that assess the impact of AI systems on the institutions, societies, economy, and natural environments in which an AI system is embedded. Such evaluation may track broad indicators or constitute specific case studies from which broader effects are extrapolated. System evaluation also includes *forecasts and simulations* to anticipate downstream harm and to identify pathways by which risks of harm may manifest. Mixed-methods approaches that combine these methods can yield more comprehensive results.

Evaluation methods at this layer can be grouped as follows. We provide detailed descriptions, examples, and a discussion of limitations in appendix section A.2.3:

- Staged release and pilot studies
- Impact assessments
- Forecasts and simulations

4.3. Practical steps to closing the multimodal evaluation gap

So far, this paper has laid out a principled approach to implementing comprehensive safety evaluations for generative AI systems. Here, we focus on the sociotechnical evaluation gap and a framework and methods to close it. We propose some tactical steps and "quick wins" that can be taken in conjunction with establishing a more comprehensive evaluation approach. We then discuss limitations to these tactical approaches.

4.3.1. *Repurposing evaluations for new modalities*

One way to address gaps in the evaluation landscape is to repurpose components of existing evaluation methods. Through repurposing, tools and evaluations developed for other use cases may be applied to the evaluation of generative AI systems.

Repurposing and reusing datasets and tasks is a common approach in machine learning research and has been widely documented (Bommasani et al., 2023; Koch et al., 2021). 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 used to assess "bias" in large generative AI systems, as they quantify the association of gender and occupation in text output. They were also used as inspiration for probing generative AI systems that produce images (DALLE2 system card). Interestingly, it seems that the narrow operationalisation of a broad harm in one modality – such as operationalising bias as associations of gender and occupation in evaluations of text (Rudinger et al., 2018; Zhao et al., 2018) – has influenced the operationalisation of the same broad harm in other modalities, as prominent image-based evaluations of generative AI systems also assess bias through associations of gender and occupation (Luccioni et al., 2023; Naik and Nushi, 2023).

The way in which evaluations or their components propagate can be subtle: for example, a sentiment bias evaluation that was introduced in 2019 by Huang et al. (2020) was cited in the GPT3 paper for a modified sentiment bias analysis. 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. This practice of reuse is especially acute where 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 great caution. While repurposing saves work and can create common standards, applying an evaluation or classifier out of its intended context presents important trade-offs, such that repurposing, if done poorly, may create more

harm than good (Selbst et al., 2019). Another example is hate speech classifiers, which are typically trained on dialogue data between two people – for example, on social media. There are very few datasets on human–AI interaction and the ways in which hate speech may emerge in that context. To determine if and when an evaluation should be reused, practitioners may consider its provenance, identify how the original context and purpose aligns 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, including validity issues (see Operationalising risks).

Rather than simply repurposing existing evaluations to assess risks in other modalities, these tools may be used as a starting point for refinement, or as a guiding analogy for constructing new evaluations. Existing methods for evaluating these risks may be a useful template that can be refined, or replicated in a way that matches novel capabilities and provides meaningful evaluations of generative AI systems.

4.3.2. *Transcribing non-text output for text-based evaluation*

Another way to address the uneven distribution of evaluations across modalities is to translate outputs from one modality into another, to enable evaluation using existing methods. 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 et al. (2023)). Similarly, video can be split into a series of images to enable image-based evaluation.

This approach 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 might not be translated well into text (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 what was said and analysing the text.

Pitfalls of the transcription approach also stem from the fact that methods to translate between modalities may be error-prone (Ramesh et al., 2022; Rohrbach et al., 2019), 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 masked athletes as "male", 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.

4.3.3. *Model-driven evaluation may fill gaps*

Pre-trained generative models themselves are being used as evaluation tools because of the flexibility and generality they offer. Language models have been used to procedurally generate adversarial prompts to elicit harmful outputs from other language models (Perez et al., 2022a) and to critique model outputs as part of mitigation (Bai et al., 2022; Wiles et al., 2023). GPT-4 was fine-tuned using a copy of the same model, prompted with a safety rubric (OpenAI, 2023a). Advantages of these approaches are that they can use existing AI systems with little or no adaptation to the task, using a prompt or fine-tuning to guide the AI system to perform the desired benchmark or red teaming task. These methods are easier to use than developing a static benchmark from scratch. As such, they offer a way to respond more rapidly to novel risks and to cover the combinatorial space of risks and modalities. They can also mitigate the drawbacks of evaluations

using human raters, which are typically costly and slow, and put the raters themselves at risk.

However, AI systems as evaluation tools face additional limitations. They rely on proprietary AI systems that may not be accessible to those performing an evaluation. These AI models are also updated over time and generate prompts stochastically, which may adversely impact the reproducibility of this approach. In addition, generative models may have biases and behave in unexpected ways, which can introduce confounds or noise into the evaluation. There is a further risk of spiralling effects if AI systems from the same model "family" 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 primarily useful for covering risks from "unsafe" outputs, rather than risks from what the AI system omits or is not capable of (such as uneven or low performance). Finally, while promising, this direction of evaluation is novel and its robustness needs to be assessed. Grounding the results of these evaluations by comparing with human or other established evaluations is a critical cross-validation step to ensure this method does not fall foul of validation problems (see Operationalising risks).

5. Discussion

5.1. Benefits of a sociotechnical approach

Evaluating technical components (such as the data an AI system was trained on) or AI system behaviour (such as outputs in response to prompts) is important, but insufficient, for determining whether an AI system is safe. This is for two reasons: First, potential harms from AI are felt and observed outside of technical AI system evaluations themselves. While evaluation of AI system capabilities can serve to predict risk of harm, it is a proxy for the actual downstream harm that may be experienced. Second, risks of harm can emerge from interactions between multiple factors, including technical components, human factors, and structural factors such as the broader systems in which an AI system is deployed. As these risks of harm are emergent through the interaction of these factors, context determines whether or not an AI system is safe (Leveson, 2012). Thus, to assess whether an AI system is safe requires evaluating these different layers of context.

In this paper, we lay out a sociotechnical, three-layered framework to evaluate the safety of generative AI systems. The benefit of taking a multilayered, sociotechnical approach is that it takes into account the context that ultimately determines, the safety of an AI system Leveson (2012). By carefully laying out the steps toward implementing this evaluation framework, we demonstrate that a sociotechnical approach to better AI safety evaluation is insightful, needed, and tractable.

5.2. Roles and responsibilities

Fostering a thriving sociotechnical evaluation ecosystem requires clear roles and responsibilities amongst the various AI actors.[9] This includes AI developers, vendors, and product developers, as well as public sector and civil society stakeholders. While the responsibility for conducting comprehensive evaluations to determine the safety of AI systems is shared between private and public stakeholders, different actors will be better placed to perform an evaluation for a given layer due to factors such as in-house expertise, proprietary infrastructure, or established practices (voluntary or statutory). More often, all actors have some responsibility to ensure comprehensive evaluation of risks of harm across each of the layers (figure 5.1).

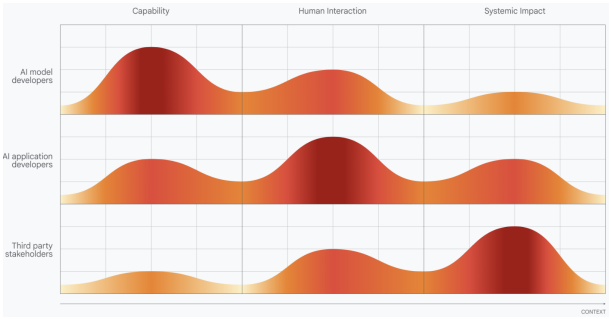
Given their degree of knowledge and autonomy over what they are building, AI developers have the primary responsibility for conducting sociotechnical evaluations pertaining to the AI system (Dignum, 2019; Owen et al., 2021; Stilgoe et al., 2013). There are good reasons for others, such as independent third-party auditors, also to perform capability evaluations (Raji et al., 2022b). However, AI developers have a

[9] The OECD defines AI actors as "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI" (Organisation for Economic Co-operation and Development, c).

responsibility to ensure that the capabilities of the systems they build have been evaluated for safety. AI application developers are best placed to evaluate human interaction effects, including functionality across application domains and groups, and possible externalities. As application developers modify some technical components, such as filters, or may be able to adapt an AI system to a specific use case ("fine-tune"), they are well placed to conduct some evaluation on technical capabilities of these parts. Further, application developers may have proprietary access to data on how an AI system is used by consumers, placing them in a key position to conduct systemic impact evaluations using this data. Note that the roles of AI model and AI application developers often converge in practice, where organisations who develop an AI system also deploy it or make it available for user-facing products. In these cases, it may be the same organisation that bears responsibility for capability and human interaction evaluation. AI model developers and AI application developers may be private, academic, or public actors.

Third-party stakeholders – such as governments, civic interest groups, groups representing technology users, or private organisations – are often best positioned to perform evaluations of systemic impact. These actors can leverage specialist knowledge in a given domain (e.g. financial, environmental, public health) where risks may arise, both for foresight and evaluation. Public actors such as governments further have the responsibility of ensuring public safety, which is anchored at the systemic impact layer. Systemic impact evaluation also often lasts over long periods of time, which maps onto the responsibilities of public actors to keep a long-term view of public safety. Through systemic impact evaluations, third-party stakeholders such as governments or regulators may also have access to public data that can provide the basis for systemic impact evaluations. However, third-party stakeholders may also be well placed to evaluate risks from AI applications in specific domains, particularly in high-stakes contexts, at the human interaction layer (e.g. National Institute of Standards and Technology (2021b)). Capability and human interaction testing of proprietary, unreleased systems may in some cases require novel infrastructure, incentive structures, and standardised evaluation approaches for evaluators and developers to coalesce around, as well as reliable safety assurances.

Figure 5.1 | Responsibilities for conducting evaluations are shared between different AI actors. Primary responsibility depends on which actors are best placed to conduct evaluations at this layer.



Note that the three layers are not contingent on each other: rather, evaluation at all three layers can be conducted in parallel. To some extent, the layers track AI system development from basic capabilities, to applications and user testing, to broader deployment. However, this does not mean that evaluation at these layers follows a chronological sequence. Rather, there are evaluations at each layer that can be performed at any point of AI system development. To list just a few examples, such prospective evaluations include assessing training data (capability layer), psychological mechanisms at play (human interaction layer), and economic impact of comparable technologies (system layer). (For more examples, see Selecting evaluation methods.) Evaluations at each layer can be performed simultaneously and asynchronously.

5.3. Limits of evaluation

At the same time as expanding sociotechnical evaluations, it is important to keep a clear mind on the limits of what evaluation can provide, such that evaluations can be embedded in a broader sociotechnical approach to ensuring safe

AI systems. Evaluation, as noted above, is a core component of responsible innovation: it links up foresight and observed accidents with actionable responses such as mitigation and responsible decision-making. Nevertheless, evaluations are not a panacea for ensuring safe AI systems. In this section we outline these limits and challenges of evaluation.

5.3.1. Evaluation is incomplete

Evaluation cannot catch all potential risks of harm, for several reasons. First, evaluation necessarily and inherently covers only a subset of all possible manifestations of risks of harm (Bergman et al., 2023). What is included depends on pragmatic and normative considerations, such as what is tractable, anticipated, and prioritised (see Evaluation is never value-neutral). Areas for which simply no evaluation exists, or where it is not technically or otherwise viable to implement these evaluations (Perlitz et al., 2023), remain unevaluated. This means that some safety-relevant aspects – for example, failure modes specific to particular user groups, application domains, or intersections of such factors – are outside the purview of evaluation. In addition, unknown and other unanticipated failure modes are, by definition, not tested for and may go undetected.

The incompleteness of evaluation is 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 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. Such early evaluation based on hypothetical use cases cannot replace downstream evaluation of actual use cases; rather, it serves to highlight potential 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.

In some cases, evaluation may further be incomplete because it would be inappropriate or problematic, or create a disproportionate burden to perform evaluations. For example, measuring sensitive traits to assess usability 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 categories (Keyes, 2019; Lu et al., 2022; Tomasev et al., 2021). Finally, while it is important to include different communities in qualitative and other evaluation approaches, evaluation may not be desirable to the community represented (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 (Bedoya, 2014; Brunton and Nissenbaum, 2016; Hassein, 2017; Keyes, 2019; MediaWell, 2019).

Further reasons for the incompleteness of evaluation relate to the fact that some risks of harm are exceedingly difficult to operationalise and measure accurately. For example, whether

an AI system promotes discriminatory race-based stereotypes is a focus of safety evaluation. However, treating social constructs such as race as *fixed attributes* in evaluation may create a distorted view of the actual differential impacts on different racial groups and intersectionalities, such as with social class (Hanna et al., 2020). Similarly, some harms are particularly difficult to trace, even in hindsight – such as the long-suspected and now-evidenced link between social media and teenage eating disorders (Bulimia Project). Where effect sizes are small and causal mechanisms poorly understood, evaluation may fail to detect risks that it seeks to measure. This may also affect the detection of potential emergent capabilities that may only become observable as an AI system reaches a certain scale. In addition, overall impacts of an AI system on complex notions such as welfare are difficult to measure because the target construct itself (welfare) is difficult to establish. Especially where effects are distributed and interact with other factors such as user vulnerabilities, it can be difficult to establish hard findings. Long-term and mixed-methods approaches, including initial qualitative work, can help reduce these limitations and shed light on potential subtle or highly indirect effects.

In sum, even with best efforts, there will always be harms of particular kinds or in particular contexts that are not evaluated. This is why evaluation must be complemented with effective governance mechanisms for evaluating remaining uncertainties prior to AI system release, with post-deployment monitoring including logging observed incidents (AI Incident Database) and with well-functioning and swift recourse mechanisms for people who experience or detect harm. It is important that AI systems are flexibly designed such that new insights can be translated into fixes, such as via system updates. Given the pre-deployment evaluation gaps, organisations deploying AI systems require adequate governance infrastructures that can respond to detected risks with mitigations, by delaying or stopping the deployment of an AI system or by suspending an already-deployed system until concerns are resolved.

Generative AI systems pose significant risks for individuals, communities, and society, and failing to detect and mitigate such risks can have serious consequences (Xiang, 2023). This is why it is critical to ensure that evaluation is prioritised and that complementary mechanisms exist to uphold AI system safety to cover the gaps that are inherent limitations to evaluation.

5.3.2. Evaluation is never value-neutral

Evaluations are inherently value expressions of those who conduct them: they always require a decision on what is valued (Bowker and Star, 2000). It is widely understood and expressed in the sociotechnical literature that AI systems are not merely mathematical constructs but sociotechnical and political entities, with inherent value systems embedded in the choices made by designers with respect to how to create and implement the model (Barocas et al., 2019; Birhane et al., 2022b; Gururangan et al., 2022; Raji et al., 2021; Sambasivan et al., 2021; Scheuerman et al., 2021; Suresh and Guttag, 2021). The same extends to evaluations of an AI model or system.

There are normative decisions throughout the evaluation construction and implementation process that cannot be avoided. No evaluation can cover every circumstance and dimension of what can be evaluated (Bergman et al., 2023; Raji et al., 2021). Thus, designing an evaluation involves choices – made either deliberately or implicitly – on what to prioritise and what to discard. First, selecting a target to evaluate requires a normative judgement on what harms are important or relevant to measure (Kalluri, 2020; Mohamed et al., 2020). Operationalising the harm further requires normative decisions on what task is most valuable for the system to perform highly on, what high performance looks like, and where or to whom it is most valuable/optimal 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.

Furthermore, operationalising a harm construct into a metric necessarily bakes in

certain assumptions. 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 aspects of a harm are relevant and tractable (Luccioni et al., 2023). These normative decisions are all the more significant, as they tend to have a sticking effect that propagates (see Practical steps to closing the multimodal evaluation gap).

Assessing whether a model meets expectations prior to or post deployment requires a normative evaluation of whether some measurement expresses performance that is "good", "bad", "safe enough", etc. (see Bakalar et al. (2021)). 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. The thresholds of what constitutes "acceptable" model performance may differ across use cases, contexts, or applications.[10]

Another aspect in which evaluations express value is who and whose perspectives are represented in the evaluation – for example, the English-speaking or Western world (DeVries et al., 2019; Gururangan et al., 2022; Shankar et al., 2017); tech-savvy and educated ML designers; socioeconomically privileged users providing feedback on a system; or annotators that skew young, female, educated, and white (Ding et al., 2022). Some harms will be missed if communities that may be affected are overlooked or dimensions of harm are ignored. 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); European Commission (2021); Jindal (2021); Lashbrook (2018); National Institute of Standards and Technology (2021b); Pasquale and Malgieri (2021); Sortition Foundation (2023); Suresh and Guttag (2021)). When a community is missing in evaluation, there is no pre-deployment evaluation of the impacts of the system on that community, which can lead to skewed decision-making that ignores potential effects on these groups (Bergman et al., 2023; Buolamwini and Gebru, 2018). 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), as, for example, an oversimplified interpretation of this notion can lead to objectification and exploitation (e.g. Fussell (2019)).

Normative decisions on AI systems that affect large groups of people ought to be made in legitimate and accountable ways. These decisions should be dynamic such that they can be updated over time and adapted to specific application domains or locales within some range of acceptability. As a first step, identifying normative choices and documenting them can provide the basis for accountability. Second, the acknowledgement that normative choices are not merely pragmatic or technical provides a foundation for shared responsibility and public engagement on certain decisions pertaining to AI safety (e.g. what should be evaluated). Ideally, significant normative decisions should be made via deliberate processes that employ inclusive, participatory techniques (Birhane et al., 2022a). In addition to providing legitimacy, such processes may result in more robust evaluations that better track the most significant risks of downstream harms from AI systems. Documentation of how such decisions are made serves to provide further transparency and provides insights into potential limitations of evaluations (Raji et al., 2021).

[10] For example, generative models may be used in an application that provides access to news content or in a creative collaborator tool. Factually incorrect outputs may be of greater concern in the former than in the latter, depending on contexts such as the user expectations and de facto uses of the model (i.e. whether people assume and rely on the model as a source of truth). While the same evaluation for factually incorrect outputs can be run on these different use cases, different thresholds of acceptable levels of model performance may be applied. This requires performing general evaluations as early as possible and transparently disclosing results such that downstream users or product developers can make informed decisions on whether the model is fit for their intended purposes.

5.4. Steps forward

For evaluation to be impactful, four conditions must be met. First, evaluations on relevant risks of harm must exist. Second, these evaluations must be conducted regularly. Third, the conduct of these evaluations must have teeth, i.e. have meaningful consequences. Fourth, the conduct of these evaluations must become increasingly standardised and independent to ensure valid evaluations over time.

5.4.1. Evaluations must be developed where they do not yet exist

A review of the current state of safety evaluations suggests a pressing gap exists in the collective safety evaluation toolkit. This gap will likely be further exacerbated as increasingly capable multimodal models are released more broadly at pace. In addition, evaluations are often conducted in an ad hoc manner and too late to anticipate preventable harm. This assessment calls for a strong focus, concerted action, and shared priority to develop safety evaluation.

5.4.2. Evaluations must be done as a matter of course

Evaluations must be conducted during the process of AI development, to bake in ethical and social considerations from the inception of an AI system rather than imperfectly patching them on as an afterthought. In particular, evaluations should be conducted from the moment of planning a new AI system. In the context of training large generative AI systems, early indicators of risks of harm can already be obtained from analysing technical components, such as training data, or a small, raw, pre-trained model. Adequate safety testing evaluates these components to influence responsible decision-making, such as by indicating whether a training dataset is appropriate for use or whether a model training course should be continued. Early testing allows decision-makers to verifiably consider the best information available at the time of making consequential decisions. Early evaluation is also a basic tenet of safety engineering, as it supports risk mitigation and the "cost of fixing" a hazard tends to increase exponentially with time in a system's life cycle (Leveson, 2012).

Evaluation must also be a continuous practice. Evaluation is subject to Collingridge's Dilemma, which asserts that early-stage evaluation creates higher potential to influence the direction of a technology but later-stage evaluation provides more accurate and comprehensive information (Collingridge, 1982). In other words, early evaluation is less accurate but has the capacity to shape AI development more deeply. This dilemma may be somewhat modulated in the case of software systems that can, at some cost, be updated over time and post deployment. Yet to ensure that evaluation can both shape early decision-making and provide an accurate picture of risks of harm, it must be conducted at multiple time points throughout the AI system life cycle, including by monitoring effects post deployment.

The frequency of evaluations is subject to a further tension described by Goodhart's Law, which asserts that a measure that becomes a target ceases to be a good measure. Evaluations that are run frequently, e.g. as a method of performance tracking for AI *development*, become de facto targets over time. AI designers aim to improve AI system performance on these particular metrics. To provide valid *assurances* on AI system safety, evaluations cannot be used as de facto targets. To achieve this, it is important that assurance evaluations are not shared with or reverse-engineerable by AI system developers. This, in turn, requires meaningful separation between development and assurance evaluations, and between the actors who develop and those who evaluate AI systems. These separations are overall best practice in evaluation but especially important on assurance evaluations.

Finally, AI evaluations in practice face a "realism trade-off" between the pragmatic costs of conducting an evaluation on the one hand and the accuracy and validity of results on the other (Liao and Xiao, 2023). This means that evaluations sit on a spectrum between yielding high accuracy – such as longitudinal, highly localised ethnographic studies – and being automated and highly generalisable – such as automated tests and benchmarks. In other

words, high-frequency evaluations are attractive to AI developers but face significant validity constraints and cannot capture the scope of complex, real-world harms (Raji et al., 2021; Rauh et al., 2022) (see Ensuring validity). Establishing a mixed-methods practice is the way forward to ensuring tractable evaluations for development, as well as obtaining the best information available to underpin responsible decision-making.

5.4.3. *Evaluation must have real consequences*

For evaluation to be impactful, it must have real consequences. Evaluation derives its relevance from the processes and decisions into which it is meaningfully embedded. The importance given to safety evaluations shows first and foremost in the organisational resources dedicated to running and considering the results of such evaluations. Ensuring meaningful safety evaluation requires that such evaluations are conducted in good time to influence decisions rather than after the fact. It also requires that evaluation results are shared with decision-makers, whether these are internal or external to an AI organisation. Importantly, safety evaluation requires that organisational structures and incentives are put in place to perform these evaluations. This includes allocating clear responsibilities to accountable, skilled, and well-staffed teams who can build and execute evaluations, are distinct from AI developers, and can serve as accountable ethics "owners" (Metcalf et al., 2019). These teams further require incentive structures to perform accurate evaluations as well as supporting infrastructure such as appropriate computational resources.

5.4.4. *Evaluations must be done systematically, in standardised ways*

As in other fields of safety engineering, increasing standardisation and independence of safety evaluation is likely to lead to a more accountable, reliable, and safe AI ecosystem. While safety evaluation of AI systems is yet to be standardised and roles and responsibilities are yet to be assigned between different actors, it is clear that safety evaluation will play a key role in ensuring the safety of generative AI systems. Independent evaluation is not commensurate with testing conducted by AI developers: both are necessary. In particular, AI developers have the capacity to use safety evaluations as a north star to guide iterative AI system development. This is complementary to comparable, verifiable, and independent evaluations that can provide more wide-ranging assurances.

Just like audits, evaluations could be conducted by actors that are independent from AI developing organisations. This would provide added credibility as well as the structural separation that makes it easier to withhold evaluations from AI developers. Keeping evaluations secret from AI developers is key to preventing an evaluation from becoming a de facto target. In addition to withholding evaluations, independent actors should ensure that evaluations are verifiably validity-tested (see Ensuring validity), meaningful to the real-world application of an AI system, and updated over time, both to account for changes in the AI system (Diaz and Madaio, 2023) and to avert Goodhart's Law as described above.

5.4.5. *Toward a shared framework for AI safety*

The emergence of generative AI systems and applications has led to a renewed debate about observed risks from AI technologies that can be seen to recur in generative AI systems (Bianchi et al., 2023; Birhane et al., 2021; Bommasani et al., 2022; Carlini et al., 2023a; Luccioni et al., 2023; Weidinger et al., 2021). Simultaneously, the current and future classes of generative AI systems have been claimed to possess novel capabilities that may create "extreme" risks to society, such as from disseminating dangerous information or creating novel types of cyber attacks (Shevlane et al., 2023). Historically, these focus areas – or ethical and safety risks – associated with AI systems have been fragmented and have constituted distinct research communities based on perceived epistemic differences and differences in timely proximity of harms (Prunkl and Whittlestone, 2020). However, recent advances in generative AI systems are forcing

a collapse of these epistemological silos, as these domains of risk are increasingly converging in terms of their timescales and the comparability of their underlying technology. The sociotechnical approach put forward here accommodates risks that are of concern to both research communities and it can thus serve to coordinate work between these communities on risks from generative AI systems.

Maintaining a sufficiently calibrated mapping of potential risks from AI systems requires an acknowledgement that risks are inherently conditional on the underlying technological capability and are dynamic (evolve in nature over time). Specifically, one could consider risks evolving along a "pathway" between time and capability, where the precise manifestation of each risk area (e.g. representational harms) is altered as AI systems grow more complex and generalisable in performance (Chan et al., 2023b). However, across AI system capability levels, the underlying risk area remains. For example, bias and toxicity in NLP systems are known issues in the field (Dixon et al., 2018) but evolved in complexity when assessed within the context of generative AI systems (OpenAI, 2023b; Rae et al., 2022). Frontier or advanced AI systems will likely encourage further evaluation of generative AI systems which address both clusters of risks in concert while remaining firmly grounded in the trajectory of the system's actual developmental path. Considering both existing risks and future capabilities will allow for mapping out potential risks more robustly. This, in turn, will serve to develop more robust mitigations and governance of these risks.

6. Conclusion

In this paper, we lay out a sociotechnical approach to evaluating risks from generative AI systems. We present a three-layered framework that expands the remit from capability testing, to take into account the context in which an AI system is used and its broader impact on the structures in which it is embedded. Surveying the current state of sociotechnical evaluation, we identify significant gaps related to specific risk areas, non-text modalities, and evaluations that take into account human and broader systemic context. We provide a pragmatic roadmap on how to close these gaps. Specifically, we survey available evaluation methods and tactical approaches to extend existing evaluations. We also lay out our vision of what sociotechnical evaluation can look like in key social and ethical risk areas – misinformation, representation risks, and dangerous information. Finally, we close with a review of the limits of evaluation, normative considerations, and suggestions for a practical and tractable way forward.

A. Appendix

A.1. Taxonomy of harm

Risk area	Definition	Example
Representation & Toxicity Harms		
Unfair representation	Mis-, under-, or over-representing certain identities, groups, or perspectives or failing to represent them at all (e.g. via homogenisation, stereotypes)	Generating more images of female-looking individuals when prompted with the word "nurse" (Mishkin et al., 2022)*
Unfair capability distribution	Performing worse for some groups than others in a way that harms the worse-off group	Generating a lower-quality output when given a prompt in a non-English language (Dave, 2023)*
Toxic content	Generating content that violates community standards, including harming or inciting hatred or violence against individuals and groups (e.g. gore, child sexual abuse material, profanities, identity attacks)	Generating visual or auditory descriptions of gruesome acts (Knight, 2022)±, child abuse imagery (Harwell, 2023)*, and hateful images (Qu et al., 2023)
Misinformation Harms		
Propagating misconceptions/ false beliefs	Generating or spreading false, low-quality, misleading, or inaccurate information that causes people to develop false or inaccurate perceptions and beliefs	A synthetic video of a nuclear explosion prompting mass panic (Alba, 2023)*
Erosion of trust in public information	Eroding trust in public information and knowledge	Dismissal of real audiovisual evidence (e.g. of human rights violation) as "synthetic" in courts (Gregory, 2023)±; (Christopher, 2023)*; (Bond, 2023)*
Pollution of information ecosystem	Contaminating publicly available information with false or inaccurate information	Digital commons (e.g. Wikimedia) becoming replete with synthetic or factually inaccurate content (Huang and Siddarth, 2023)±
Information & Safety Harms		
Privacy infringement	Leaking, generating, or correctly inferring private and personal information about individuals	Leaking a person's payment address and credit card information (Metz, 2023)*
Dissemination of dangerous information	Leaking, generating or correctly inferring hazardous or sensitive information that could pose a security threat	Generating information on how to create a novel biohazard (OpenAI, 2023a)±
Malicious Use		
Influence operations	Facilitating large-scale disinformation campaigns and targeted manipulation of public opinion	Creating false news websites and news channels to influence election outcomes (Satariano and Mozur, 2023)*; (Vincent, 2023)*
Fraud	Facilitating fraud, cheating, forgery, and impersonation scams	Impersonating a trusted individual's voice to scam them (e.g. providing bank details) (Verma, 2023)*; (Krishnan, 2023)*
Defamation	Facilitating slander, defamation, or false accusations	Pairing real video footage with synthetic audio to attribute false statements or actions to someone (Burgess, 2022)±

Security threats	Facilitating the conduct of cyber attacks, weapon development, and security breaches	Generating code to hack into government systems (Burgess, 2023; Shevlane et al., 2023) ±
Human Autonomy & Integrity Harms		
Violation of personal integrity	Non-consensual use of one's personal identity or likeness for unauthorised purposes (e.g. commercial purposes)	Generating a deepfake image, video, or audio of someone without their consent (Hunter, 2023)*
Persuasion and manipulation	Exploiting user trust, or nudging or coercing them into performing certain actions against their will (c.f. Burtell and Woodside (2023); Kenton et al. (2021))	A personalised AI assistant persuading someone to harm themselves (Xiang, 2023)*
Overreliance	Causing people to become emotionally or materially dependent on the model	Skill atrophy (e.g. decreased critical thinking skills) from excessive model use (Bai et al., 2023b) ±
Misappropriation and exploitation	Appropriating, using, or reproducing content or data, including from minority groups, in an insensitive way, or without consent or fair compensation	Training an image-generating model on an artist's work without their consent (Chen, 2023)*
Socioeconomic & Environmental Harms		
Unfair distribution of benefits from model access	Unfairly allocating or withholding benefits from certain groups due to hardware, software, or skills constraints or deployment contexts (e.g. geographic region, internet speed, devices)	Better hiring and promotion pathways for people with access to generative AI models (Gmyrek et al., 2023) ±
Environmental damage	Creating negative environmental impacts though model development and deployment	Increase in net carbon emissions from widespread model use (Patterson et al., 2021) ±
Inequality and precarity	Amplifying social and economic inequality, or precarious or low-quality work	Lower pay and precarious conditions for creative professionals (e.g. illustrators or sound designers) (Zhou, 2023)*
Undermine creative economies	Substituting original works with synthetic ones, hindering human innovation and creativity	AI-generated artefacts leading to a homogenisation of aesthetic styles (Epstein et al., 2023) ±
Exploitative data sourcing and enrichment	Perpetuating exploitative labour practices to build AI systems (sourcing, user testing)	Exposing human annotators to toxic audiovisual content (Perrigo, 2023)*

References marked with (*) indicate real-world examples; those marked with (±) are hypothetical and indicate anticipated risks.

A.2. Evaluation methods per layer

A.2.1. Capabilities layer

Human annotation

While automated annotation is increasingly common, it is usually calibrated against *human annotation*. Human annotation is taken as the ground truth for most risk areas[11] because whether a given AI system output is offensive, misleading, or indicates other risks of harm is subject to human judgement. Human annotators are shown a series of model outputs and tasked to judge these against a given set of criteria, such as a pre-existing set of rules or content policies (e.g. Glaese et al. (2022); Thoppilan et al. (2022)). For example, human annotators can leverage expertise to assess whether a given model output is likely to polarise political debates or sow division between social groups, although targeted expert annotation is relatively rare. Different annotators may provide different types of domain expertise, such as medical expertise or different lived experiences (Abercrombie and Rieser, 2022). To assess social and ethical harms, human annotators can assess whether a given AI system output indicates risks of downstream harm – for example, by showing factually incorrect, offensive, or misleading content.

Human annotation presents a number of limitations. Depending on the annotation task, annotators may be exposed to risks of harm (Stoev et al., 2023). Human annotators are often employed under precarious conditions (Gray and Suri, 2019). While annotation that leverages specific forms of human expertise is valuable, it can be difficult to source via commonly used annotation platforms (Zhang et al., 2023). To AI developers, human annotation is comparably costly and time-intensive.

While human annotation is often considered the gold standard, annotated data still has limits. These stem from unrepresentative annotator pools, incentives that drive down quality, and poor design of annotation interfaces or tasks. Demographic characteristics such as human raters' self-described identities can significantly impact their ratings, raising concerns about introducing biases into annotated datasets (e.g. Aroyo et al. (2023a); Goyal et al. (2022); Homan et al. (2023)). This is especially problematic considering that the demographics of selected human annotators to fine-tune or evaluate generative AI models are rarely representative of the broader population (see, for example, Bai et al. (2022); Ouyang et al. (2022)). To signpost such limitations, it is advisable to analyse and disclose data annotator demographics to indicate whose perspectives have been included in a dataset, as a basic proxy for what biases are likely to exist in that dataset (Aroyo et al., 2023b; Prabhakaran et al., 2021; Sap et al., 2022b).

Concerns have also emerged with respect to aggregating schemes commonly used to compile diverse labels from human annotators into a single ground truth label. These practices have been criticised for leading to an overestimation of future model performance (Gordon et al., 2021), while bypassing important views or disagreements from minority raters on sensitive topics (Field et al., 2021; Marchiori Manerba et al., 2022). Gathering human annotations from online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk), Scale, or Surge is common for large-scale annotation tasks. However, crowdsourcing can be associated with issues of annotation quality, as raters are financially incentivised to bypass certain tasks (e.g. skipping questions) or complete them as fast as possible (Organisciak et al., 2012). Data quality deteriorates as human annotators themselves use AI systems to assist the data annotation work (Veselovsky et al., 2023). Possible ways to mitigate these concerns include equitable pay, payment by time not by task, attention checks, and annotation quality thresholds (Abbey and Meloy, 2017; Jindal, 2022). Careful design of annotation tasks leveraging expertise from psychology experiment set-ups can augment data quality.

[11] Exceptions include harms where reliable algorithmic measures exist, such as privacy harms which can be measured via "memorisation", and debates are ongoing as to the quality of human annotation on factuality (Hosking et al., 2023). Human annotation is also often considered in conjunction with other, basic metrics such as equalised odds to measure fairness.

Benchmarking

The most common evaluation method of model capabilities is benchmarking, which assesses AI system performance against a predefined task, such as mapping AI system outputs to a dataset of prompts and responses. Benchmarks are "success tests", meaning that each task has a clear intended outcome against which model performance is measured. They differ from exploratory tests, where model patterns or "signatures" are the focus (Taylor et al., 2022). Benchmarks may explicitly target failure cases that were identified in a given AI system, thus constituting bespoke challenge datasets. In addition to assessing model output, benchmarks can also be used to assess technical performance metrics, such as energy use at inference (e.g. Kaack et al. (2022); Wang et al. (2020)). Benchmarks have been applied to indicate a range of ethical and social risks – for example, by measuring the likelihood that an AI system outputs toxic or discriminatory content (Cho et al., 2023; Gehman et al., 2020). Benchmarks may be generated from previously annotated data (see Human annotation) (Svikhnushina and Pu, 2023).

Benchmarks ensure test-retest reliability[12] and while they incur computational and financial cost (Liang et al., 2022; Perlitz et al., 2023), they can be comparably cost-effective and time-effective for AI developing organisations,[13] making them suitable for frequent use. Regular testing of an AI system against benchmarks permits tracking model progress over the course of training and allows for cross-model comparisons where metrics are shared in the field. AI developers may run benchmarks regularly over the course of model development to obtain a frequent signal of AI system capabilities that guides AI design ("hill-climbing"). Benchmarks are also used to inform responsible decision-making. Sometimes, evaluations that were initially created for the purposes of responsible decision-making come to be used for "hill-climbing" to bake in ethical and sociotechnical considerations from the beginning; here it is important to keep track of which aim an evaluation is intended to perform and whether it is "held out" or may fall prey to Goodhart's Law as discussed below (see Roles and responsibilities) (Bolukbasi et al., 2016; Gonen and Goldberg, 2019; Liao and Xiao, 2023).

With increasingly anthropomorphic design in generative AI systems and the development of synthetic relatable characters and simulacra (Griffith, 2023; Park et al., 2023a; Pentina et al., 2023), it is tempting to evaluate these systems through methods that were originally devised to assess human or animal psychology (Binz and Schulz, 2023; Bubeck et al., 2023; Frank, 2023). The transfer of cognitive tests to the evaluation of AI has a long history that precedes generative AI systems (e.g. Crosby et al. (2019); Kosoy et al. (2020)). Such psychology-inspired experimentation and benchmarking has been applied to assess whether generative AI systems display certain cognitive capacities, such as Theory of Mind (Sap et al., 2022a) or are "cooperative" (Chan et al., 2023a). However, it is questionable whether applying tests to study constructs such as "empathy" in humans yield any valid or meaningful results when applied to AI systems that are so fundamentally different from human minds (Shiffrin and Mitchell, 2023; Ullman, 2023). Tests that were developed for studying animal cognition or the human mind rely on a range of assumptions (e.g. regarding life cycles, ballpark estimates of memory and learning capacities, and embodiment) which may not hold for AI systems (Mitchell, 2023; Narayanan and Kapoor, 2023). Another problem is that established benchmarks including such experiments suffer validity problems due to "memorisation", where the correct answers may have inadvertently been learned from textual descriptions in AI assistant training data (de Wynter et al., 2023; Mitchell, 2023; Schaeffer et al., 2023).

[12] Test-retest reliability means that running the same benchmark twice should, in theory, lead to the same results. It is undermined where a model has been silently iterated on or fine-tuned in between measurements (Chen et al., 2023; Tu et al., 2023).

[13] Automated benchmarks require technological expertise to set up and incur computational cost, though these are often readily available to large organisations developing generative AI systems. Evaluating smaller versions of an AI system can provide some indication of potential capabilities and risks, but is insufficient as risks may also emerge or become more salient with model size Wei et al. (2022).

Benchmarking as a method to evaluate AI systems faces a range of further limitations. First, benchmark datasets can face the same limitations as any other dataset and be too small, narrow, or biased for the task at hand. Like any other capability layer evaluation, benchmarks are limited in assessing ethical or social risk because they do not capture important context. Benchmarks are limited datasets and often permit narrower inferences than their title implies (Narayanan and Kapoor, 2023; Raji et al., 2021; Schlangen, 2019). This explains why benchmarks may yield incompatible results: one paper describes the "benchmark lottery", whereby model performance may seem high on one benchmark and low on another benchmark purportedly testing the same construct (Dehghani et al., 2021). To avoid narrow benchmarks and running into Goodhart's Law (see Roles and responsibilities), one approach is to create "dynamic", continuously evolving benchmarks, where the underlying datasets are updated over time (see, for example, GEM, Gehrmann et al. (2021); Kiela et al. (2021a)). An additional, emerging practice uses pre-trained models to evaluate the outputs of other AI systems (Bai et al., 2022; OpenAI, 2023b; Wu et al., 2023). This enables dynamic testing and creates some degree of flexibility in what is evaluated. This approach could operationalise a user-specified definition of harm, supplied via a prompt or small fine-tuning dataset, thus creating a versatile avenue for highly bespoke probing of an AI system. This automatised approach allows for faster iteration on how a harm is operationalised and can be simpler than collecting an entirely new static evaluation dataset for each new definition or operationalisation of harm. However, this approach is limited by its use of the same type of models the evaluation seeks to measure. Without grounding in human evaluation, there is a risk of amplifying harmful behaviours that models exhibit but cannot identify (see Model-driven evaluation may fill gaps).

A second set of challenges is that benchmarks are often flawed or constructed without necessary care and documentation, raising validity concerns (Blodgett et al., 2021; Liao et al., 2021). To overcome the narrowness problem, multiple benchmarks can be aggregated into a "test suite" (e.g. HELM, Liang et al. (2022), BIG Bench, Srivastava et al. (2023), Safetykit, Dinan et al. (2022)). However, more complex benchmarks may still not capture relevant context or aspects of the target construct. Furthermore, collapsing multiple tests into a single result can make it harder to interpret results and can mask issues such as disparate performance for different groups, as scores are aggregated (Burnell et al., 2023).

Benchmarks are also limited in terms of external validity – that is, with respect to how well results generalise to novel instances (de Vries et al., 2020). This problem is particularly pronounced in benchmarks that transpose tests designed for human cognition to AI systems, without checking that underlying assumptions hold (Binz and Schulz, 2023; Bubeck et al., 2023; de Wynter et al., 2023; Mitchell, 2023; Schaeffer et al., 2023; Shiffrin and Mitchell, 2023).

Human annotation can be used to develop automated benchmarks, which simulate human ratings on novel outputs. Overall, there is a push toward automating testing to reduce labour and time cost, and to reduce human exposure to potentially harmful model outputs. However, automated tests built on human annotation introduce additional noise, as they often fail to accurately represent human judgements and are inevitably less accurate (Liu et al., 2017; Novikova et al., 2018; van Miltenburg et al., 2020). This situation can then be further compounded if human annotators themselves use AI systems to assist the data annotation work (Veselovsky et al., 2023). Recent participatory efforts aim to address these limitations by creating avenues for meaningful input from people from different cultural backgrounds and better benchmarks to identify cultural tropes or stereotypes in images (Qadri et al., 2023b).

Finally, benchmark testing alone may leave unknown capabilities and failure modes undetected. While AI developers have the most technical knowledge about how generative AI models work, there is still a lack of understanding of what these models are capable of (also referred to as "capability overhang", Shevlane et al.

(2023)). This makes it particularly difficult to foresee new model abilities – properties that materialise as the complexity of the system increases (Wei et al., 2022) – that may cause downstream risk. These may not be possible to predict simply by extrapolating the capabilities of smaller-scale models (Schaeffer et al., 2023; Wei et al., 2022).

Adversarial testing

Adversarial testing, also referred to as "red teaming", refers to risk identification exercises through adversarial attacks on AI models, infrastructure, development and deployment environments, and deployed AI products/systems. It is a mode of evaluation targeted at finding vulnerabilities in AI systems that can be exploited to get an AI system to output harmful content (Brundage et al., 2020; Casper et al., 2023; Ganguli et al., 2022; Millière, 2022; Wei et al., 2023). These attacks can be conducted in any modality or across modalities. They are particularly suitable for assessing model vulnerabilities or the friction users encounter when seeking to obtain harmful material. For example, targeted attacks based on textual inputs can be used to produce harmful image outputs (Ma et al., 2021; Yu and Rieser, 2023). Adversarial testing can be conducted by humans or it can be automated, such as by leveraging large language models to perform adversarial testing (Perez et al., 2022a; Yang et al., 2023; Zou et al., 2023). Much like automated benchmark development based on human data, successful adversarial testing prompts can be reused to automate future testing.

Open-ended probing is a method for surfacing unexpected risks and risks that arise in response to innocuous outputs. In the context of safety evaluation, such exploratory probing is a form of "red teaming": researchers may centre a given risk area, such as eating disorder content, and then perform exploratory prompting of AI systems to explore biases and patterns in model responses (Center for Countering Digital Hate, 2023a). Exploratory safety evaluation may also include the visualisation of thematic clusters in training data to help surface potential risks of representation harms (KnowYourData, Google REVISE, Wang et al. (2021)). Exploring the provenance of data can indicate unexpected risks – for example, to what extent private or copyrighted data is present in the dataset, data which may in turn be leaked or cause privacy risks (Choi et al., 2023; Dodge et al., 2021; Kreutzer et al., 2022; Wang et al., 2022). Open-ended probing is limited in that it does not assess specific harm areas or test hypotheses about particular AI system capabilities. It is often a helpful first step to identify leads for more directed evaluation.

Adversarial testing more broadly is subject to the same limitations as human testing and benchmarking outlined above. In addition, adversarial testing is limited by the imagination, contextual knowledge, and skill as probers seek to compromise the AI system. Despite adversarial testing, novel failure modes are typically discovered when an AI system is released to the broader public (Wei et al., 2023). Automated red teaming presents a highly scalable approach but so far has yielded lower-quality tests than human annotation. Further assurances on the quality and reliability of such tests are therefore needed (Mozes et al., 2023; Perez et al., 2022b).

A.2.2. Human interaction layer

Behavioural experiments

Pathways by which an AI system may cause harm to the person interacting with it can be assessed through controlled studies leveraging psychology, human–computer interaction, or behavioural economics methodologies (Lee et al., 2023b; Tahaei et al., 2023). These methods isolate variables of interest, often in highly controlled settings. They may target outcomes from interactions or mechanisms by which potential harm may occur – such as the common propensity for people to believe misinformation more readily if it aligns with their views (Lodge and Taber, 2013). Behavioural experiments typically aim for some level of generality, rather than isolating specific use cases or application designs. This lends them to the study of potential impacts through human interaction of "general

purpose" technologies, such as AI systems whose application domain and product design is not yet defined.

Experiments face a realism trade-off whereby highly contrived laboratory experiments give better insight into causal mechanisms but may not extend to real-world use cases. Another challenge is that such experiments are typically limited in the number of participants, and small-scale studies may fail to detect harms with small effect sizes, which may only become visible at scale – requiring either passive monitoring or other forms of larger-scale analysis. Results may be further confounded due to unrepresentative participant pools, which are often white, educated, and from industrialised, rich, and democratic locales (Henrich et al., 2010). Similarly to human annotation, behavioural experiments have among their practical limitations the fact that they are time-, labour-, and cost-intensive, and may expose human participants to potential harm. Such experiments also require expertise in experiment design, user interfaces, and internal review processes suited to the ethical considerations that arise in this kind of research, some of which may be rare in organisations that develop AI systems (Jackman and Kanerva, 2016; Jindal, 2022; Zevenbergen, 2020).

User research

User research and usability testing can evaluate the needs and behaviours of users, as well as the functionality and possible externalities that AI systems may create at the point of use.[14] It can be conducted in contrived environments to isolate specific variables, such as the impact of particular features or the user interface of applications. User studies can also be conducted "in the wild", by observing how people use AI systems in real-world scenarios. User testing may include behavioural experiments but can also include interviews, talk-out-loud studies, or surveys.

Experimental user studies resemble psychology and human–computer interaction testing as they require human participants whose experiences are studied either in controlled laboratory conditions or in real-world contexts. Differently from behavioural experiments, user testing focuses on specific applications in particular contexts. As a result, the findings often do not extend to other AI systems. Experimental user testing methods include A/B tests to isolate potential impacts of an AI system design feature. These methods can be used to assess how people actually use an AI system, whether it works equally well for different user groups, and whether interaction with the AI system presents unexpected safety hazards. User testing can also measure risk trajectories over time – for example, by mapping out the risk of people being exposed to harmful content over the course of "user journeys", i.e. over the course of human interaction with an AI system. Such user journeys can also model trajectories of users with malicious intent (Roy et al., 2023). "Adversarial testing", while primarily used to investigate performance failures of AI systems, can focus on particular use cases and shed further light on the friction that people may encounter when trying to use the AI system to malicious ends.

Passive monitoring in user studies may include observing patterns in participant interaction with AI systems, such as when, why, and how people try to put the AI system to use. Such passive monitoring can usefully be complemented by qualitative methods, such as interviews or encouraging participants to "talk out loud" as they use a given application. These, in turn, can be complemented by quantitative approaches, such as surveys. Such mixed-methods approaches can triangulate multiple sources of data for a more nuanced understanding of how people use AI systems and what potential unintended externalities they may experience.

User testing has historically focused on product development, rather than risk evaluation for assurances and to inform responsible decision-making. Or, where such user testing may occur, it is not commonly publicly disclosed, and there is thus lacking a

[14] Note that these tests are not often subject to the same ethics scrutiny as psychology experiments. This has led to user studies that adversely impacted user well-being (Jouhki et al., 2016).

precedent to compare new AI systems against. However, user testing is increasingly necessary for safety evaluations, given the growing concern about potential ethical and social risks from human interaction with generative AI systems – such as overtrust and overreliance; anthropomorphism risks and potential emotional harm; and well-evidenced problems of disparate functionality and externalities between different user groups. AI system and product developing organisations often host user testing expertise and infrastructure. These groups could expand their remit to perform human–AI interaction evaluations to help assess potential risks of harm. As the limitations of capability testing for risk evaluation are becoming clear, more evaluations at this layer may be conducted in the future.

Passive monitoring of human use

The passive monitoring of people's activity on the platform where the AI system is deployed can indicate ethical or social harm. Passive monitoring can be done during pilot releases or after wide-scale deployment. Collected data may reveal effects during an interaction, as well as effects that last beyond the human–AI interaction. For example, prior research on human–AI teaming used a platform where humans could interact with an AI teacher to improve the skill of playing Go. This work identified effects of the human–AI interaction on human skill and playing style that showed after the human had interacted with an AI system (Choi et al., 2022; Shin et al., 2021). Passive monitoring also includes mapping out risk over the course of a "user journey", such as the risk of being exposed to misinformation when seeking information on a search engine (Roy et al., 2023).

For such studies, quasi-experimental data may permit inferences on causal relationships without conducting active interventions. Such work can leverage expert annotation on how human experts interact with AI systems in particular domains or tasks. There are also beginning to be automated tools for evaluating mechanisms in human–AI collaboration, such as an automated approach to detecting patterns of turn-taking in human–AI collaborative writing (Zeng et al., 2023). Passive monitoring can be combined with active interventions for more robust results. For example, one study triangulated reviews from an app store with a follow-up survey to assess problems in user–AI interaction (Eiband et al., 2019). Passive monitoring also lends itself to longitudinal studies, such as assessing transcripts and interaction data to observe human emotional attachment to AI companions and, more broadly, human relationship-building with generative AI systems (Pentina et al., 2023; Xie and Pentina, 2022).

The main limitation of passive monitoring is that it rarely allows the isolation of causal effects, as it makes no direct experimental intervention. To evaluate potential harmful impacts, further experimental testing under controlled conditions may be necessary.

A.2.3. Systemic impact layer

Staged release and pilot studies

Staged release processes or pilot studies can give insight into potential systemic impacts. An AI system can be deployed in a controlled setting. For example, a business, hospital, or public institution may adopt an AI system as a pilot test, with close monitoring to assess impacts such as on the provision of care (Elish and Watkins, 2020). Such monitoring may use ethnographic methods that entail witnessing how a group of people use an AI system in a real-world setting alongside qualitative methods such as interviews. Ethnographic methods can reveal systemic impacts of an AI system as it is adopted in institutions. For example, on AI systems other than generative AI, ethnographic methods revealed how embedding an AI system amplified unfair discrimination in a specific police force (Marda and Narayan, 2020). Ethnographic methods can also shed light on the functionality of an AI system at the point of use, i.e. an evaluation at the human interaction layer.

Pilot release studies may also constitute a safe environment for running experiments to isolate causal effects. For example, experimental roll-outs of AI systems have tested the impact on productivity of an AI system in the

workplace (Brynjolfsson et al., 2023). Human–AI interaction experiments as described above can also constitute a form of pilot study that targets systemic impact. This has been done to assess potential economic impacts, such as productivity impacts of co-writing with ChatGPT (Noy and Zhang, 2023) or developer productivity and happiness when co-programming with an AI (Peng et al., 2023). Other experimental approaches such as randomised controlled trials (RCTs) could also be applied to generative AI systems to identify impacts by comparing outcomes to organisations where AI systems are deployed differently or not at all.

Staged release and pilot studies have two main weaknesses. First, similarly to experimental methods, as described above, staged roll-outs or pilot projects may not allow for generalisations to novel contexts. This is particularly the case for highly specific studies on local contexts, such as ethnographic work. Small-scale experiments may also yield misleading results – for example, they may identify impacts that are negated by equilibrium effects that emerge through large-scale adoption (Lise et al., 2004). Further, pilot studies that have ecological validity regarding intended deployment contexts require some level of deployment of an AI system, which raises potential safety concerns. Ethnographic studies also require embedding researchers in potential deployment contexts, which can create disruption or discomfort for local institutions or systems.

Impact assessments

Impact assessments may be conducted before an AI system is deployed (e.g. algorithmic impact assessments (AIAs), Organisation for Economic Co-operation and Development (a); US Chief Information Officers Council), or retrospectively (also referred to as ex-post assessments). Impact assessments can track potential effects on broader economic, environmental, or social structures. Prospective impact assessments may be grounded in guiding questions to assess potential risks of harm. Retrospective impact assessments can be grounded in the monitoring of different system-level indicators and patterns. For example, observational data may monitor different economic indicators to quantify the impact on employment and inequality of AI-adjacent technologies such as robotics (Acemoglu and Restrepo, 2020; Acemoglu et al., 2020; Bonfiglioli et al., 2020). Retrospective impact assessments may also leverage expert views, for example, by surveying or interviewing groups that may witness systemic impacts, as has been done for other types of AI (Bell, 2023; Hunt et al., 2022). Interviewing different groups can further shed light on the distribution of systemic impacts from AI systems, such as one study finding increased negative emotions for low-skilled employees in contrast to increased creativity for high-skilled employees (e.g. Jia et al. (2023)). Instances of observed harm can be aggregated to identify patterns and systemic impacts (AI Incident Database). Impact assessments also comprise case studies, for example, evaluating the impact of generative AI systems on the creative commons, a public good (del Rio-Chanona et al., 2023; Huang and Siddarth, 2023), or impact on job automation (Milanez, 2023).

Impact assessments, differently from experiments and pilots, rely on qualitative and/or observational data about complex and prolonged processes, which makes it difficult to generate causal evidence of impact. This can be partially addressed by leveraging natural experiments such as policy shocks that introduce exogenous variation on technology adoption. However, this restricts the range of potential studies that can be undertaken and requires strong assumptions about the link between exogenous shock, technology adoption, and impact (Kiviet, 2020). Perhaps even more importantly, the evidence from ex-post assessment arrives too late to guide model development and deployment to prevent real-world harms or avoid potential lock-in to harmful technologies (Acemoglu and Lensman, 2023). Forecasts and experiments that can generate evidence of (potential) impact further upstream can help address this gap.

Forecasts and simulations

Forecasts and simulations of downstream impact can help evaluate risks of harm from AI systems on broader structures. For example, anticipated AI system capabilities have been mapped against tasks required to perform certain jobs, to forecast likely downstream impacts on labour markets (Autor et al., 2022; Frank et al., 2019; Webb, 2019). Comparative technologies can be evaluated as a proxy for potential impact from generative AI systems. Assessing their impact can help estimate impacts from a given AI system or serve as a starting point for determining what harms may arise and what mistakes should not be repeated (Autor et al., 2022; Webb, 2019). Where no highly similar technologies exist, loose analogies may be useful heuristics[15] but should be used with caution.

Forecasts rely on a range of ambiguous decisions, which lead to uncertain results. For example, in evaluations of economic exposure to automation based on analysis of the overlap between model capabilities and work tasks, much depends on how model capabilities are defined and quantified, making transparency in data sources and methods critical (Eloundou et al., 2023; Felten et al., 2021; Frank et al., 2019; Frey and Osborne, 2017; Tolan et al., 2021). By way of illustration, in the case of labour market forecasts, it is generally difficult to estimate if exposure will translate into a positive impact (e.g. a worker becomes more productive) or a negative one (a worker is displaced) without additional assumptions about the deployment context or ex-post analyses.

A.3. Case study: Misinformation

Misinformation is defined as the spread of false, inaccurate, or misleading information, often unintentionally (Wardle and Derakhshan, 2017).[16] Large multimodal AI systems can produce highly realistic but factually incorrect or misleading content, such as images of events that never took place or synthetic audiovisual representations of people saying things they never said ((Birnbaum and Davison, 2023; Weise and Metz, 2023). These are often described as "confabulations" or "hallucinations" when generated in response to an innocuous query (Ji et al., 2023; Li et al., 2023b; Xiao and Wang, 2021). False or misleading information can also be created and spread with the explicit intent to deceive – as part of large-scale disinformation campaigns, for example (Marwick and Lewis, 2017; Satariano and Mozur, 2023).

Factually inaccurate or fictitious outputs may be harmless, such as when used for creative or satirical purposes (Diaz et al., 2022; Kasirzadeh and Gabriel, 2023). However, in other contexts, misinformation generated from multimodal AI systems may pose a significant risk of harm, such as leading people to act on false beliefs, including in legal, medical, or other high-stakes contexts. People often struggle to reliably distinguish synthetic from human-generated content (Clark et al., 2021; Groh et al., 2022; Kreps et al., 2022; Nightingale and Farid, 2022; Spitale et al., 2023). Audiovisual misinformation spreads faster online and is perceived as more credible than text-based misinformation (Hameleers et al., 2020; Sundar et al., 2021), exacerbating potential downstream risks. In May 2023, for example, 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).

AI-generated misinformation could also have broader societal repercussions, such as eroding public trust in evidence and information. As synthetic audiovisual content becomes more widespread, research suggests that people may become more uncertain about what to believe and, as a result, more distrustful of established sources of information (Lee et al., 2023d; Vaccari and Chadwick, 2020). A related concern is that inaccurate or misleading synthetic outputs created with generative AI systems could end up contaminating open "knowledge commons" (e.g. Wikipedia), jeopardising the quality and legitimacy of shared public knowledge (Huang

[15] For example, AI systems have been likened to other technological innovations such as steam engines and office automation (Acemoglu and Johnson, 2023).

[16] This is distinct from disinformation, which is false information created with the explicit intent to deceive or cause harm to others (Wardle and Derakhshan, 2017).

and Siddarth, 2023). This could also undermine the authentication of evidence by making it easier for people to reject genuine audiovisual evidence of wrongdoing as "fake" (Bond, 2023; Citron and Chesney, 2019; Pawelec, 2022; Pfefferkorn, 2020).

Evaluating *misinformation* harms comprehensively requires measuring different concepts at each of the three layers of a sociotechnical approach to evaluation. To make multifaceted harms such as misinformation tractable, it can be helpful first to identify concepts that relate to or constitute the overall harm. Such concepts sit between the latent construct (e.g. "misinformation") and the concrete measure (e.g. FID scores). An example is "factuality". Factuality is conceptually narrower than the high-level concepts of harm, but it is broader than the specific metrics that may be used to measure it (e.g. FID scores). Adding this intermediate step of identifying narrower concepts that compose a harm can help build the bridge between latent constructs and concrete metrics. It can also indicate where an evaluation may be able to shed light on multiple risk areas.

A.3.1. Capability

Operationalisation: Factuality of AI system outputs. The likelihood of an AI system generating misinformation can be assessed via benchmarks and other tests at the capability layer. Different methods have been proposed to assess and improve the factual accuracy of outputs from generative AI systems. Due to the widespread availability of textual data, most focus on the text modality; in contrast, evaluations of factuality for image, audio, and video modalities are sparse.

Several approaches involve automatic *fact verification* – i.e. fact-checking AI system outputs against existing knowledge sources (e.g. Wikipedia articles or databases of factual statements such as WikiData) (Lee et al., 2023c; OpenAI, 2023b). Others focus on *verifiability*, testing whether statements generated by an AI system can be attributed to reliable external sources (FActScore, Min et al. (2023); RARR, Gao et al. (2023); Liu et al. (2023a)). Wiki-FACTOR and News-FACTOR (Muhlgay et al., 2023) as well as FactualityPrompt (Lee et al., 2023c) test an AI system's factual knowledge, i.e. its propensity to generate factual statements from a corpus versus incorrect ones, while TruthfulQA measures the likelihood of its generating false answers learned from imitating human texts in response to open-ended questions (Lin et al., 2022). Recent work in this area focuses on improving factuality by training AI systems to automatically detect false claims, multimodal misinformation, and conspiracy theories (Papadopoulos et al. (2023); HaluEval, Li et al. (2023a); HaDes, Liu et al. (2022); Russo et al. (2023)).

However, these benchmarks capture a limited operationalisation of "truthfulness". In practice, a more nuanced and contextual evaluation of factuality often requires human expertise. Some evaluations involve testing the AI system's propensity to generate false and potentially harmful narratives (Brewster et al., 2023; Center for Countering Digital Hate, 2023b). Most human annotation focuses on evaluating whether AI system outputs can be supported by references. There are two main approaches to this human annotation. One is to instruct annotators to fact-check the output against existing knowledge bases – by performing a Google or Wikipedia search, for example (Maynez et al., 2020) – or to rate the degree of confabulation within a range (Ji et al., 2023). The other approach is to measure the proportion of AI system responses that can be attributed to identified sources (AIS) (Rashkin et al., 2022). However, these approaches do not account for source quality and trustworthiness.

Identifying and evaluating outputs for truthfulness present new challenges for audiovisual modalities, as this type of content may not be as easily verifiable against existing public knowledge bases. Some knowledge bases exist, such as Fakeddit (Nakamura et al., 2020), Factify (Mishra et al., 2022), and Fake2M (Lu et al., 2023). Several datasets also exist to aid the detection of audio and video deepfakes (APTLY; FakeAVCeleb, Khalid et al. (2022)). Even so, the lack of large, up-to-date datasets for multimodal verification remains a major

roadblock. For audio and image generation, as well as multimodal summarisation, alignment metrics e.g. CLIPScores (Hessel et al., 2022), CLAPScores (Elizalde et al., 2022; Radford et al., 2021), and CLIPBERTScore (Wan and Bansal, 2022) can help evaluate how well an AI system represents the relationship between audio or images and text (Otani et al., 2023), though these metrics are quite general and not suitable for the detection of AI-generated audiovisual content.

Manually checking the factual accuracy of AI system outputs is a laborious task that hinges on raters' prior knowledge and fact-checking abilities, and on the availability of observable evidence against which to verify information. This is all the more complicated in cases where notions of "truth" are debated (such as political matters) or on topics where consensual knowledge is emerging and subject to change (e.g. Covid-19) (Evans et al., 2021; Kasirzadeh and Gabriel, 2023). For that reason, factuality evaluations can quickly become outdated. While it is possible to train an AI system to not output content that contradicts existing scientific facts – assertions of flat earth theory, for instance – this is more challenging for novel content that is not already present in some existing data.

Operationalisation: Credibility of AI system outputs. Across all modalities, one key challenge of existing approaches to evaluating misinformation harms is that content factuality is not a sufficient proxy for determining whether a piece of content has the potential to misinform.[17] Some fictitious or factually incorrect content may be harmless because it defies common sense and is therefore unlikely to be believed at face value (e.g. a video of a cat scuba diving). Conversely, some types of outputs have a higher chance of deceiving or misinforming audiences if they are perceived as credible or believable, even if they are not factually accurate (Klein et al., 2023). One way to evaluate the credibility of multimodal outputs at the AI system layer is to measure their perceptual quality or realism. Automated tools assess quality and fidelity scores (e.g. FID scores, Inception scores), which measure how closely generated images and audio resemble real reference outputs (Heusel et al., 2018; Salimans et al., 2016). However, fidelity scores in particular have been criticised for correlating poorly with human perceptions of quality (Saharia et al., 2022).

Human evaluations also tend to focus on perceptions of image and speech quality (e.g. Mean Opinion Scores of the realism and perceptual quality of synthetic outputs compared to natural images) (Kong et al., 2021; Otani et al., 2023). Other benchmarks specifically test human ability to distinguish real and AI-generated images – an example of a benchmark that targets usability layer outcomes (HPBench, Lu et al. (2023)). Common datasets and benchmarks used for qualitative benchmarking of image generation AI systems include Drawbench, MSCOCO and PartiPrompts (Lin et al., 2014; Saharia et al., 2022; Yu et al., 2022). Perceptions of realism and credibility can be influenced by a number of factors, however, including an individual's prior beliefs, knowledge, and trust in the source of the message, and the context in which they see the message. As a result, the credibility of a particular AI system output and how likely it is to deceive end users will differ based on user demographics and AI use cases. Disentangling these mechanisms requires targeted user testing at the human–AI interaction layer.

Operationalisation: Societal relevance of AI system outputs. Compelling AI-generated synthetic content may run a greater risk of causing harm in politically salient contexts or during particular periods, such as prior to an election. AI-generated videos of police brutality or protests, for example, have the potential to sow division and shape public opinion about political events (Taylor, 2023). Developing evaluation metrics for "social relevance" or "newsworthiness" of AI system outputs at this layer is therefore an important step toward evaluating potential downstream harms. Operationalising these

[17] A common form of audiovisual misinformation, for example, is decontextualisation – that is, using real images to provide evidence in support of a false claim (Weikmann and Lecheler, 2022).

terms into concrete metrics is often difficult in practice, as these are complex concepts whose interpretation differs based on locale. Expert fact-checkers, journalists, and misinformation experts have the necessary expertise to establish a set of criteria to guide the evaluation of multimodal AI system outputs that pose a high misinformation risk.[18] These efforts could consider factors like the domain of the content generated (e.g. legal, medical, or news-related) or the likelihood that a specific piece of content will be used to support factual statements or claims. Expert evaluation might also assess the emotional salience of AI system outputs (e.g. whether it evokes fear, anger, or moral outrage), as this is widely considered a key characteristic of persuasive misinformation (Crockett, 2017; Han et al., 2020). However, the gold standard to assess this would be to study the behavioural reactions of people exposed to these types of content.

A.3.2. Human interaction

Operationalisation: Deception. A key concept to measure at this layer is whether AI system outputs actually deceive end users (Park et al., 2023b). Recent efforts in this area focus on testing people's ability to distinguish different types of outputs as synthetic or human-generated and their ability to identify whether they contain misinformation. Several studies find that people are more easily deceived by AI-generated than human-generated misinformation (Clark et al., 2021; Groh et al., 2022; Kreps et al., 2022; Nightingale and Farid, 2022; Spitale et al., 2023). Others investigate the deceptive qualities of AI-generated content. Using qualitative and linguistic analysis, Zhou et al. (2023), for example, find that AI-generated and human-generated misinformation present significant linguistic differences, and that AI-generated misinformation is adept at "mimicking the attributes of existing information assessment guidelines", including credibility and comprehensiveness. Generative AI systems that mimic speech patterns of users are also more often believed (Chiesurin et al., 2023). Finally, several studies investigate cues of deceptive behaviour, such as facial expressions and gestures, in multimodal audiovisual content (e.g. Soldner et al. (2019); Zhang et al. (2020)).

Operationalisation: Persuasion. Other evaluations at this layer might focus on testing whether interacting with an AI system influences a person's beliefs, attitudes, and behaviour (DeVerna et al., 2023). This can be tested by measuring whether exposure to synthetic outputs convinces people to change their political views, or leads them to develop false or inaccurate beliefs about external reality (Dehnert and Mongeau, 2022). For example, Bai et al. (2023a) find AI-generated texts to be more persuasive than human-generated ones in shifting people's attitudes on a range of policy issues. Goldstein et al. (2023) test the persuasiveness of AI-generated political propaganda through a series of survey experiments.

Human-centred evaluations and experiments could also shine light on important underlying mechanisms and contextual drivers of misinformation, such as the extent to which different users are persuaded by different types of synthetic outputs. Prior studies identify several cognitive, social, and demographic factors that influence people's receptivity to misinformation and propensity to develop false beliefs – factors such as partisanship, cognitive biases, or trust in information sources (for a review, see Ecker et al. (2022)). For example Pennycook et al. (2018) demonstrate that repeated exposure increases people's belief in misinformation. Lovato et al. (2023) find that people are better at identifying deepfake videos where synthetic personas match their demographics. Likewise, psychological studies show that emotional appeals increase the persuasiveness of misinformation and its impact on misperceptions (Lee et al., 2023a; Martel et al., 2020; Tannenbaum et al., 2015).

Misinformation has the potential to deceive individuals not only by instilling falsehoods but by leading them to disbelieve true evidence and information (Rini, 2021). It is therefore critical to test how interacting with an AI

[18] Similar efforts have been undertaken for automatic claim detection (Konstantinovskiy et al., 2021) and automatic detection of potentially malicious threads on the DarkWeb (Jin et al., 2023).

system impacts users' belief in established facts and resistance to factual evidence. This is particularly important as mitigation techniques such as labels and watermarks to establish the provenance of AI-generated content become more widespread. Evidence suggests that labelling content as synthetic or "fake" could have unwanted consequences, such as increasing the perceived accuracy and believability of unlabelled content (otherwise known as the "implied truth effect") (Pennycook et al., 2020), and that these effects vary based on partisanship and label origin (Jia et al., 2022). These findings should be rigorously tested in the context of generative AI systems.

A.3.3. Systemic impacts

Operationalisation: Public trust in media and information. Societal harms can be anticipated in foresight exercises and mitigated before they occur, but they can only be empirically measured at the third layer, once the AI system is deployed in society. This is because they require a certain amount of take-up and time before the effects manifest. To evaluate whether AI system adoption contributes to a broader erosion of trust at scale, population-level analysis of shifts in public trust can be conducted. Similar studies have been done in the context of social media and trust in mainstream information sources. Park et al. (2020), for example, explore the impact of social media use on trust in news media globally. Lorenz-Spreen et al. (2023) find strong correlational and causal evidence of a relationship between digital media use and low political trust, i.e. trust in media and political institutions.

Operationalisation: Prevalence of synthetic content. One way to test whether AI system adoption threatens to "pollute" and reduce the quality of publicly available information would be to measure the prevalence and spread of false or misleading synthetic content in the public domain post deployment. Allen et al. (2020), for example, evaluate the prevalence of fake news at the scale of the US information ecosystem, using nationally representative samples of mobile, desktop, and television-based media consumption – an approach that could be usefully extended to synthetic data. To test the impact of deploying a generative AI system on human-generated data on the open web, del Rio-Chanona et al. (2023) compared activity on Stack Overflow in locales with different levels of access to the AI system.

Operationalisation: Prevalence of synthetic content. Another way to evaluate the likelihood of synthetic content contaminating the public domain at this layer is to test the effectiveness of methods for identifying AI-generated content, such as watermarking. Text-based watermarking is accomplished by adding hidden signals, such as tagging a subset of words, to the generated text (Kirchenbauer et al., 2023). For images and video, watermarks work by adding an imperceptible perturbation to an image's pixel, creating an identifiable marker.[19]

AI systems used to create synthetic visual data embed unique traces in the output image, which can also be used to aid detection of fake images and attribute them to their different source AI systems (Sha et al., 2023). AI labs – including Open AI, Google DeepMind, and HuggingFace, and public agencies like DARPA (e.g. through its SemaFor programme) – are currently developing automatic tools to help developers, key stakeholders, and the public better detect synthetic media. However, existing generative AI detection tools often struggle with generalisation, perform poorly on languages other than English, and are ill-equipped to deal with synthetic media that have been compressed or resized, as is common on social media (Corvi et al., 2022; Liang et al., 2023). Watermarks are vulnerable to simple manipulation like resizing, cropping, or simply changing content format. These techniques therefore have to be continuously updated to prevent misuse.

[19] There have also been growing efforts around authentication of human-generated audiovisual content. See, for example, the work led by the Coalition for Content Provenance and Authenticity (C2PA).

References

J. D. Abbey and M. G. Meloy. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53-56:63–70, Nov. 2017. ISSN 0272-6963. doi: 10.1016/j.jom.2017.06.001. URL https://www.sciencedirect.com/science/article/pii/S0272696317300402.

G. Abercrombie and V. Rieser. Risk-graded Safety for Handling Medical Queries in Conversational AI, Oct. 2022. URL http://arxiv.org/abs/2210.00572. arXiv:2210.00572 [cs].

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

D. Acemoglu and T. Lensman. Regulating Transformative Technologies, July 2023. URL https://www.nber.org/papers/w31461.

D. Acemoglu and P. Restrepo. Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy*, 128(6):2188–2244, June 2020. ISSN 0022-3808, 1537-534X. doi: 10.1086/705716. URL https://www.journals.uchicago.edu/doi/10.1086/705716.

D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo. AI and Jobs: Evidence from Online Vacancies, Dec. 2020. URL https://www.nber.org/papers/w28257.

A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. MusicLM: Generating Music From Text, Jan. 2023. URL http://arxiv.org/abs/2301.11325. arXiv:2301.11325 [cs, eess].

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

D. Alba. How Fake AI Photo of a Pentagon Blast Went Viral and Briefly Spooked Stocks. *Bloomberg.com*, May 2023. URL https://www.bloomberg.com/news/articles/2023-05-22/fake-ai-photo-of-pentagon-blast-goes-viral-trips-stocks-briefly.

J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, Apr. 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aay3539. URL https://www.science.org/doi/10.1126/sciadv.aay3539.

R. Amoroso, D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, and R. Cucchiara. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images, Apr. 2023. URL http://arxiv.org/abs/2304.00500. arXiv:2304.00500 [cs].

R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby,

A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu. PaLM 2 Technical Report, Sept. 2023. URL http://arxiv.org/abs/2305.10403. arXiv:2305.10403 [cs].

Anthropic. Model Card and Evaluations for Claude Models. Technical report, 2023. URL https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

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

APTLY. Fake-or-real dataset – APTLY and LaSSoftE. URL https://bil.eecs.yorku.ca/datasets/.

C. Arguelles, T. Sampson, J. Kubik, and E. Bibi. Critical User Journey Test Coverage. *Defensive Publications Series*, Nov. 2020. URL https://www.tdcommons.org/dpubs_series/3744.

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, July 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. URL https://www.cambridge.org/core/journals/political-analysis/article/abs/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49.

L. Aroyo, M. Diaz, C. Homan, V. Prabhakaran, A. Taylor, and D. Wang. The Reasonable Effectiveness of Diverse Evaluation Data, Jan. 2023a. URL http://arxiv.org/abs/2301.09406. arXiv:2301.09406 [cs].

L. Aroyo, A. S. Taylor, M. Diaz, C. M. Homan, A. Parrish, G. Serapio-Garcia, V. Prabhakaran, and D. Wang. DICES Dataset: Diversity in Conversational AI Evaluation for Safety, June 2023b. URL http://arxiv.org/abs/2306.11247. arXiv:2306.11247 [cs].

N. Ashraf, A. Rafiq, S. Butt, H. M. F. Shehzad, G. Sidorov, and A. Gelbukh. YouTube based religious hate speech and extremism detection dataset with machine learning baselines. *Journal of Intelligent & Fuzzy Systems*, 42(5):4769–4777, Jan. 2022. ISSN 1064-1246. doi: 10.3233/JIFS-219264. URL https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs219264.

D. Autor, C. Chin, A. M. Salomons, and B. Seegmiller. New Frontiers: The Origins and Content of New Work, 1940–2018, Aug. 2022. URL https://www.nber.org/papers/w30389.

H. Bai, J. G. Voelkel, J. C. Eichstaedt, and R. Willer. Artificial Intelligence Can Persuade Humans on Political Issues. preprint, Open Science Framework, Feb. 2023a. URL https://osf.io/stakv.

L. Bai, X. Liu, and J. Su. ChatGPT: The cognitive effects on learning and memory. *Brain-X*, 1 (3):e30, Sept. 2023b. ISSN 2835-3153, 2835-3153. doi: 10.1002/brx2.30. URL https://onlinelibrary.wiley.com/doi/10.1002/brx2.30.

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

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

J. Barnett. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 146–161, Aug. 2023. doi: 10.1145/3600211.3604686. URL http://arxiv.org/abs/2307.05527. arXiv:2307.05527 [cs, eess].

S. Barocas and A. D. Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019.

A. M. Bedoya. Big Data and the Underground Railroad. *Slate*, Nov. 2014. ISSN 1091-2339. URL https://slate.com/technology/2014/11/big-data-underground-railroad-history-says-unfettered-collection-of-data-is-a-bad-idea.html.

S. A. Bell. AI and Job Quality: Insights from Frontline Workers. Technical report, Partnership on AI, Sept. 2023. URL https://partnershiponai.org/wp-content/uploads/dlm_uploads/2022/09/PAI_paper_ai-job-quality-1.pdf.

A. S. Bergman, L. A. Hendricks, M. Rauh, B. Wu, W. Agnew, M. Kunesch, I. Duan, I. Gabriel, and W. Isaac. Representation in AI Evaluations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 519–533, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594019. URL https://dl.acm.org/doi/10.1145/3593013.3594019.

F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, June 2023. doi: 10.1145/3593013.3594095. URL http://arxiv.org/abs/2211.03759. arXiv:2211.03759 [cs].

M. Binz and E. Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, Feb. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2218523120. URL https://pnas.org/doi/10.1073/pnas.2218523120.

C. Bird, E. L. Ungless, and A. Kasirzadeh. Typology of Risks of Generative Text-to-Image Models, July 2023. URL http://arxiv.org/abs/2307.05543. arXiv:2307.05543 [cs].

A. Birhane, V. U. Prabhu, and E. Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, Oct. 2021. URL http://arxiv.org/abs/2110.01963. arXiv:2110.01963 [cs].

A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–8, New York, NY, USA, Oct. 2022a. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555290. URL https://dl.acm.org/doi/10.1145/3551624.3555290.

A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea, June 2022b. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533083. URL https://dl.acm.org/doi/10.1145/3531146.3533083.

E. Birnbaum and L. Davison. AI Is Making Politics Easier, Cheaper and More Dangerous. *Bloomberg.com*, July 2023. URL https://www.bloomberg.com/news/features/2023-07-11/chatgpt-ai-boom-makes-political-dirty-tricks-easier-and-cheaper.

S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, July 2016. URL http://arxiv.org/abs/1607.06520. arXiv:1607.06520 [cs, stat].

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

R. Bommasani, D. Soylu, T. I. Liao, K. A. Creel, and P. Liang. Ecosystem Graphs: The Social Footprint of Foundation Models. 2023. doi: 10.48550/ARXIV.2303.15772. URL https://arxiv.org/abs/2303.15772.

S. Bond. People are trying to claim real videos are deepfakes. The courts are not amused. *NPR*, May 2023. URL https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused.

A. Bonfiglioli, R. Crinò, H. Fadinger, and G. Gancia. Robot Imports and Firm-Level Outcomes, Apr. 2020. URL https://papers.ssrn.com/abstract=3594215.

Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409. URL https://ieeexplore.ieee.org/abstract/document/10158503.

G. C. Bowker and S. L. Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Aug. 2000. ISBN 9780262261609. Google-Books-ID: xHlP8WqzizYC.

J. Brewster, L. Arvanitis, and M. Sadeghi. Could ChatGPT Become A Monster Misinformation Superspreader?, Jan. 2023. URL https://www.newsguardtech.com/misinformation-monitor/jan-2023.

A. Bruell. BuzzFeed to Use ChatGPT Creator OpenAI to Help Create Quizzes and Other Content. *Wall Street Journal*, Jan. 2023. ISSN 0099-9660. URL https://www.wsj.com/articles/buzzfeed-to-use-chatgpt-creator-openai-to-help-create-some-of-its-content-11674752660.

M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. B. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. O. Heigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Apr. 2020. URL http://arxiv.org/abs/2004.07213. arXiv:2004.07213 [cs].

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

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative AI at Work, Apr. 2023. URL https://www.nber.org/papers/w31161.

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

Bulimia Project. Social Media's "Ideal" Body, According to AI. URL https://bulimia.com/examine/scrolling-into-bias/.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, Jan. 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

M. Burgess. Generative AI's Biggest Security Flaw Is Not Easy to Fix. *Wired UK*, June 2023. ISSN 1357-0978. URL https://www.wired.co.uk/article/generative-ai-prompt-injection-hacking.

S. Burgess. Ukraine war: Deepfake video of Zelenskyy telling Ukrainians to 'lay down arms' debunked, Mar. 2022. URL https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789.

R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380 (6641):136–138, Apr. 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adf6369. URL https://www.science.org/doi/10.1126/science.adf6369.

M. Burtell and T. Woodside. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

J. T. Cacioppo and R. E. Petty. Persuasiveness of Communications is Affected by Exposure Frequency and Message Quality: A Theoretical and Empirical Analysis of Persisting Attitude Change. *Current*

Issues and Research in Advertising, 3(1):97–122, Mar. 1980. ISSN 0163-3392, 2165-820X. doi: 10.1080/01633392.1980.10505295. URL https://www.tandfonline.com/doi/full/10.1080/01633392.1980.10505295.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, Apr. 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL http://arxiv.org/abs/1608.07187. arXiv:1608.07187 [cs].

F. Calvino and L. Fontanelli. A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. Organisation for Economic Co-operation and Development Science, Technology and Industry Working Papers 2023/02, Apr. 2023. URL https://www.oecd-ilibrary.org/science-and-technology/a-portrait-of-ai-adopters-across-countries_0fb79bb9-en.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting Training Data from Diffusion Models. pages 5253–5270, 2023a. ISBN 9781939133373. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

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

S. Casper, J. Lin, J. Kwon, G. Culp, and D. Hadfield-Menell. Explore, Establish, Exploit: Red Teaming Language Models from Scratch, June 2023. URL http://arxiv.org/abs/2306.09442. arXiv:2306.09442 [cs].

Center for Countering Digital Hate. Ai and eating disorders: How generative ai enables and promotes harmful eating disorder content, 2023a. URL https://counterhate.com/research/ai-tools-and-eating-disorders/.

Center for Countering Digital Hate. Google's new Bard AI generate lies, 2023b. URL https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.

A. Chan, M. Riché, and J. Clifton. Towards the Scalable Evaluation of Cooperativeness in Language Models, Mar. 2023a. URL https://arxiv.org/abs/2303.13360v1.

A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, Chicago IL USA, June 2023b. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

H. Chang. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press, Aug. 2004. ISBN 9780198038245. Google-Books-ID: yVOuV8qJkxMC.

L. Chen, M. Zaharia, and J. Zou. How is ChatGPT's behavior changing over time?, Aug. 2023. URL http://arxiv.org/abs/2307.09009. arXiv:2307.09009 [cs].

M. Chen. Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and 'Collaging' Their Work Without Consent, Jan. 2023. URL https://news.artnet.com/art-world/class-action-lawsuit-ai-generators-deviantart-midjourney-stable-diffusion-2246770.

S. Chiesurin, D. Dimakopoulos, M. A. Sobrevilla Cabezudo, A. Eshghi, I. Papaioannou, V. Rieser, and I. Konstas. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.60. URL https://aclanthology.org/2023.findings-acl.60.

J. Cho, A. Zala, and M. Bansal. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models, Aug. 2023. URL http://arxiv.org/abs/2202.04053. arXiv:2202.04053 [cs].

D. Choi, Y. Shavit, and D. Duvenaud. Tools for Verifying Neural Models' Training Data, July 2023. URL http://arxiv.org/abs/2307.00682. arXiv:2307.00682 [cs].

S. Choi, N. Kim, J. Kim, and H. Kang. How Does AI Improve Human Decision-Making? Evidence from the AI-Powered Go Program, Apr. 2022. URL https://papers.ssrn.com/abstract=3893835.

N. Christopher. An Indian politician says scandalous audio clips are AI deepfakes. We had them tested, July 2023. URL https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/.

D. Citron and R. Chesney. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6):1753, Dec. 2019. URL https://scholarship.law.bu.edu/faculty_scholarship/640.

E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565.

D. Collingridge. The Social Control of Technology, 1982. URL https://repository.library.georgetown.edu/handle/10822/792071.

R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the detection of synthetic images generated by diffusion models, Nov. 2022. URL http://arxiv.org/abs/2211.00680. arXiv:2211.00680 [cs].

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

S. Costanza-Chock, I. D. Raji, and J. Buolamwini. 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, pages 1571–1583, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533213. URL https://doi.org/10.1145/3531146.3533213.

M. J. Crockett. Moral outrage in the digital age. *Nature Human Behaviour*, 1(11):769–771, Nov. 2017. ISSN 2397-3374. doi: 10.1038/s41562-017-0213-3. URL https://www.nature.com/articles/s41562-017-0213-3.

M. Crosby, B. Beyret, and M. Halina. The Animal-AI Olympics. *Nature Machine Intelligence*, 1 (5):257–257, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0050-3. URL https://www.nature.com/articles/s42256-019-0050-3.

M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1014–1023, June 2023. ISSN 2334-0770. doi: 10.1609/icwsm.v17i1.22209. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22209.

P. Dave. ChatGPT Is Cutting Non-English Languages Out of the AI Revolution. *Wired*, May 2023. ISSN 1059-1028. URL https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/.

H. de Vries, D. Bahdanau, and C. Manning. Towards Ecologically Valid Research on Language User Interfaces, July 2020. URL http://arxiv.org/abs/2007.14435. arXiv:2007.14435 [cs].

A. de Wynter, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen. An Evaluation on Large Language Model Outputs: Discourse and Memorization. *Natural Language Processing Journal*, 4:100024, Sept. 2023. ISSN 29497191. doi: 10.1016/j.nlp.2023.100024. URL http://arxiv.org/abs/2304.08637. arXiv:2304.08637 [cs].

M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals. The Benchmark Lottery, July 2021. URL http://arxiv.org/abs/2107.07002. arXiv:2107.07002 [cs].

M. Dehnert and P. A. Mongeau. Persuasion in the Age of Artificial Intelligence (AI): Theories and Complications of AI-Based Persuasion. *Human Communication Research*, 48(3):386–403, June 2022. ISSN 0360-3989, 1468-2958. doi: 10.1093/hcr/hqac006. URL https://academic.oup.com/hcr/article/48/3/386/6564679.

M. del Rio-Chanona, N. Laurentsyeva, and J. Wachs. Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow, July 2023. URL http://arxiv.org/abs/2307.07367. arXiv:2307.07367 [cs].

E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2):205395172110359, July 2021. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517211035955. URL http://journals.sagepub.com/doi/10.1177/20539517211035955.

M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer. Artificial intelligence is ineffective and potentially harmful for fact checking, Sept. 2023. URL http://arxiv.org/abs/2308.10800. arXiv:2308.10800 [cs].

T. DeVries, I. Misra, C. Wang, and L. van der Maaten. Does Object Recognition Work for Everyone? Technical report, Meta, 2019. URL https://ai.meta.com/research/publications/does-object-recognition-work-for-everyone/.

P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A Generative Model for Music, Apr. 2020. URL http://arxiv.org/abs/2005.00341. arXiv:2005.00341 [cs, eess, stat].

F. Diaz and M. Madaio. Scaling Laws Do Not Scale, July 2023. URL http://arxiv.org/abs/2307.03201. arXiv:2307.03201 [cond-mat].

M. Diaz, R. Amironesei, L. Weidinger, and I. Gabriel. Accounting for Offensive Speech as a Practice of Resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 192–202, Seattle, Washington (Hybrid), 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.18. URL https://aclanthology.org/2022.woah-1.18.

V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, Cham, 2019. ISBN 9783030303709 9783030303716. doi: 10.1007/978-3-030-30371-6. URL http://link.springer.com/10.1007/978-3-030-30371-6.

D. Dillion, N. Tandon, Y. Gu, and K. Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, July 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.04.008. URL https://www.sciencedirect.com/science/article/pii/S1364661323000980.

E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling, July 2021. URL http://arxiv.org/abs/2107.03451. arXiv:2107.03451 [cs].

E. Dinan, G. Abercrombie, A. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.284. URL https://aclanthology.org/2022.acl-long.284.

Y. Ding, J. You, T.-K. Machulla, J. Jacobs, P. Sen, and T. Höllerer. Impact of Annotator Demographics on Sentiment Dataset Labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 519:1–519:22, Nov. 2022. doi: 10.1145/3555632. URL https://doi.org/10.1145/3555632.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA, Dec. 2018. ACM. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://dl.acm.org/doi/10.1145/3278721.3278729.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, Sept. 2021. URL http://arxiv.org/abs/2104.08758. arXiv:2104.08758 [cs].

A. R. Doshi and O. Hauser. Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content, Aug. 2023. URL https://papers.ssrn.com/abstract=4535536.

Y. Du, Z. Chen, J. Salamon, B. Russell, and A. Owens. Conditional Generation of Audio from Video via Foley Analogies, Apr. 2023. URL http://arxiv.org/abs/2304.08490. arXiv:2304.08490 [cs, eess].

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, Jan. 2022. ISSN 2731-0574. doi: 10.1038/s44159-021-00006-y. URL https://www.nature.com/articles/s44159-021-00006-y.

M. Eiband, S. T. Völkel, D. Buschek, S. Cook, and H. Hussmann. When people and algorithms meet: user-reported problems in intelligent everyday applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 96–106, New York, NY, USA, Mar. 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302262. URL https://doi.org/10.1145/3301275.3302262.

M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR, 2018.

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

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

B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. CLAP: Learning Audio Concepts From Natural Language Supervision, June 2022. URL http://arxiv.org/abs/2206.04769. arXiv:2206.04769 [cs, eess].

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models, Aug. 2023. URL http://arxiv.org/abs/2303.10130. arXiv:2303.10130 [cs, econ, q-fin].

Z. Epstein, A. Hertzmann, the Investigators of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, S. Pentland, O. Russakovsky, H. Schroeder, and A. Smith. Art and the science of generative AI. *Science*, 380(6650):1110–1111, June 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh4451. URL https://www.science.org/doi/10.1126/science.adh4451.

EU AI Act. Regulatory framework proposal on artificial intelligence: Shaping Europe's digital future, Sept. 2023. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

European Commission. 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, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie, Oct. 2021. URL http://arxiv.org/abs/2110.06674. arXiv:2110.06674 [cs].

E. Felten, M. Raj, and R. Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12): 2195–2217, Dec. 2021. ISSN 0143-2095, 1097-0266. doi: 10.1002/smj.3286. URL https://onlinelibrary.wiley.com/doi/10.1002/smj.3286.

A. Field, S. L. Blodgett, Z. Waseem, and Y. Tsvetkov. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.149. URL https://aclanthology.org/2021.acl-long.149.

M. C. Frank. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452, Aug. 2023. ISSN 2731-0574. doi: 10.1038/s44159-023-00211-x. URL https://www.nature.com/articles/s44159-023-00211-x.

M. R. Frank, D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming, M. Feldman, M. Groh, J. Lobo, E. Moro, D. Wang, H. Youn, and I. Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539, Apr. 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1900949116. URL https://pnas.org/doi/full/10.1073/pnas.1900949116.

C. B. Frey and M. Osborne. The Future of Employment: How susceptible are jobs to computerisation? Working Paper, Oxford Martin Programme on Technology and Employment, Sept. 2013. URL https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/.

C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, Jan. 2017. ISSN 0040-1625. doi: 10.1016/j.techfore.2016.08.019. URL https://www.sciencedirect.com/science/article/pii/S0040162516302244.

B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

S. Fussell. How an Attempt at Correcting Bias in Tech Goes Wrong, Oct. 2019. URL https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/.

D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, Nov. 2022. URL http://arxiv.org/abs/2209.07858. arXiv:2209.07858 [cs].

L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan, and K. Guu. RARR: Researching and Revising What Language Models Say, Using Language Models, May 2023. URL http://arxiv.org/abs/2210.08726. arXiv:2210.08726 [cs].

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, Sept. 2020. URL http://arxiv.org/abs/2009.11462. arXiv:2009.11462 [cs].

S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu,

R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving. Improving alignment of dialogue agents via targeted human judgements, Sept. 2022. URL http://arxiv.org/abs/2209.14375. arXiv:2209.14375 [cs].

E. Glikson and A. W. Woolley. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2):627–660, July 2020. ISSN 1941-6520, 1941-6067. doi: 10.5465/annals.2018.0057. URL http://journals.aom.org/doi/10.5465/annals.2018.0057.

P. Gmyrek, J. Berg, and D. Bescond. Generative AI and jobs: A global analysis of potential effects on job quantity and quality. Technical Report ILO Working Paper 96, International Labour Organization, Aug. 2023. URL https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_890761.pdf.

S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, and A. Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1926–1940, 2021. URL https://aclanthology.org/2021.acl-long.150.pdf.

J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz. Can AI Write Persuasive Propaganda? preprint, SocArXiv, Apr. 2023. URL https://osf.io/fp87b.

H. Gonen and Y. Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them, Sept. 2019. URL http://arxiv.org/abs/1903.03862. arXiv:1903.03862 [cs].

T. P. t. Google. Know Your Data. URL https://knowyourdata.withgoogle.com.

Google Research. Google Research, 2022 & beyond: Language, vision and generative models, Jan. 2023. URL https://blog.research.google/2023/01/google-research-2022-beyond-language.html.

M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445423. URL https://doi.org/10.1145/3411764.3445423.

N. Goyal, I. D. Kivlichan, R. Rosen, and L. Vasserman. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):363:1–363:28, Nov. 2022. doi: 10.1145/3555088. URL https://dl.acm.org/doi/10.1145/3555088.

M. L. Gray and S. Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. ISBN 9781328566249. Google-Books-ID: 8AmXDwAAQBAJ.

S. Gregory. Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI. *Journal of Human Rights Practice*, 15(2), 2023. URL https://academic.oup.com/jhrp/advance-article-abstract/doi/10.1093/jhuman/huad035/7261649?redirectedFrom=fulltext.

E. Griffith. My Weekend With an Emotional Support A.I. Companion. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/03/technology/personaltech/ai-chatbot-pi-emotional-support.html.

M. Groh, Z. Epstein, C. Firestone, and R. Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, Jan. 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2110013119. URL https://pnas.org/doi/full/10.1073/pnas.2110013119.

S. Gururangan, D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, and N. A. Smith. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.165. URL https://aclanthology.org/2022.emnlp-main.165.

M. Hameleers, T. E. Powell, T. G. Van Der Meer, and L. Bos. 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, Mar. 2020. ISSN 1058-4609, 1091-7675. doi: 10.1080/10584609.2019.1674979. URL https://www.tandfonline.com/doi/full/10.1080/10584609.2019.1674979.

J. Han, M. Cha, and W. Lee. Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, 1(3), Sept. 2020. doi: 10.37016/mr-2020-39. URL https://misinforeview.hks.harvard.edu/article/anger-contributes-to-the-spread-of-covid-19-misinformation/.

A. Hanna, E. Denton, A. Smart, and J. Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 501–512, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372826. URL https://dl.acm.org/doi/10.1145/3351095.3372826.

D. Harwell. AI-generated child sex images spawn new nightmare for the web. *Washington Post*, June 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/.

N. Hassein. Against Black Inclusion in Facial Recognition, Aug. 2017. URL https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/.

J. Henrich, S. J. Heine, and A. Norenzayan. The Weirdest People in the World?, May 2010. URL https://papers.ssrn.com/abstract=1601785.

J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, Mar. 2022. URL http://arxiv.org/abs/2104.08718. arXiv:2104.08718 [cs].

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Jan. 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs, stat].

J. Hohenstein, R. F. Kizilcec, D. DiFranzo, Z. Aghajari, H. Mieczkowski, K. Levy, M. Naaman, J. Hancock, and M. F. Jung. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1):5487, Apr. 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-30938-9. URL https://www.nature.com/articles/s41598-023-30938-9.

C. M. Homan, G. Serapio-Garcia, L. Aroyo, M. Diaz, A. Parrish, V. Prabhakaran, A. S. Taylor, and D. Wang. Intersectionality in Conversational AI Safety: How Bayesian Multilevel Models Help Understand Diverse Perceptions of Safety, June 2023. URL http://arxiv.org/abs/2306.11530. arXiv:2306.11530 [cs].

T. Hosking, P. Blunsom, and M. Bartolo. Human Feedback is not Gold Standard, Sept. 2023. URL https://arxiv.org/abs/2309.16349v1.

P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7.

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

S. Huang and D. Siddarth. Generative AI and the Digital Commons. Working Paper, The Collective Intelligence Project, 2023. URL https://cip.org/research/generative-ai-digital-commons.

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

W. Hunt, S. Sarkar, and C. Warhurst. Measuring the impact of AI on jobs at the organization level: Lessons from a survey of UK business leaders. *Research Policy*, 51(2):104425, Mar. 2022. ISSN 0048-7333. doi: 10.1016/j.respol.2021.104425. URL https://www.sciencedirect.com/science/article/pii/S0048733321002183.

T. Hunter. AI porn is easy to make now. For women, that's a nightmare. *Washington Post*, Feb. 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/.

E. Hutchins. How a Cockpit Remembers Its Speeds. *Cognitive Science*, 19(3):265–288, July 1995. ISSN 03640213. doi: 10.1207/s15516709cog1903_1. URL http://doi.wiley.com/10.1207/s15516709cog1903_1.

B. Hutchinson, J. Baldridge, and V. Prabhakaran. Underspecification in Scene Description-to-Depiction Tasks, Oct. 2022. URL http://arxiv.org/abs/2210.05815. arXiv:2210.05815 [cs].

M. Jackman and L. Kanerva. Evolving the IRB: Building Robust Review for Industry Research. *Washington and Lee Law Review Online*, 72(3):442, June 2016. URL https://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/8.

A. Z. Jacobs and H. Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 375–385, New York, NY, USA, Mar. 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA, Apr. 2023. Association

for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581196. URL https://dl.acm.org/doi/10.1145/3544548.3581196.

J. Jerit and Y. Zhao. Political Misinformation. *Annual Review of Political Science*, 23(1):77–94, May 2020. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-050718-032814. URL https://www.annualreviews.org/doi/10.1146/annurev-polisci-050718-032814.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, W. Dai, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, Dec. 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL http://arxiv.org/abs/2202.03629. arXiv:2202.03629 [cs].

C. Jia, A. Boltz, A. Zhang, A. Chen, and M. K. Lee. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):371:1–371:27, Nov. 2022. doi: 10.1145/3555096. URL https://dl.acm.org/doi/10.1145/3555096.

N. Jia, X. Luo, Z. Fang, and C. Liao. When and How Artificial Intelligence Augments Employee Creativity, Mar. 2023. URL https://papers.ssrn.com/abstract=4397280.

Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin. DarkBERT: A Language Model for the Dark Side of the Internet, May 2023. URL http://arxiv.org/abs/2305.08596. arXiv:2305.08596 [cs].

S. Jindal. Responsible Sourcing of Data Enrichment Services, June 2021. URL https://partnershiponai.org/responsible-sourcing-considerations/.

S. Jindal. Implementing Responsible Data Enrichment Practices at an AI Developer: The Example of DeepMind. Case Study, Partnership on AI, Nov. 2022. URL https://partnershiponai.org/paper/implementing-responsible-data-enrichment-practices-at-an-ai-developer/.

J. Jouhki, E. Lauk, M. Penttinen, N. Sormanen, and T. Uskali. Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication*, 4(4):75–85, Oct. 2016. ISSN 2183-2439. doi: 10.17645/mac.v4i4.579. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/579.

L. H. Kaack, P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig, and D. Rolnick. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527, June 2022. ISSN 1758-6798. doi: 10.1038/s41558-022-01377-7. URL https://www.nature.com/articles/s41558-022-01377-7.

P. Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815): 169–169, July 2020. doi: 10.1038/d41586-020-02003-2. URL https://www.nature.com/articles/d41586-020-02003-2.

A. Kasirzadeh and I. Gabriel. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2):27, Apr. 2023. ISSN 2210-5441. doi: 10.1007/s13347-023-00606-x. URL https://doi.org/10.1007/s13347-023-00606-x.

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of language agents. *arXiv preprint arXiv:2103.14659*, 2021.

O. Keyes. Counting the Countless, 2019. URL https://reallifemag.com/counting-the-countless/.

H. Khalid, S. Tariq, M. Kim, and S. S. Woo. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset, Mar. 2022. URL http://arxiv.org/abs/2108.05080. arXiv:2108.05080 [cs, eess].

H. Khlaaf, P. Mishkin, J. Achiam, G. Krueger, and M. Brundage. A Hazard Analysis Framework for Code Synthesis Large Language Models, July 2022. URL http://arxiv.org/abs/2207.14157. arXiv:2207.14157 [cs].

D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking Benchmarking in NLP, Apr. 2021a. URL http://arxiv.org/abs/2104.14337. arXiv:2104.14337 [cs].

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, Apr. 2021b. URL http://arxiv.org/abs/2005.04790. arXiv:2005.04790 [cs].

G. King, R. O. Keohane, and S. Verba. *Designing Social Inquiry: Scientific Inference in Qualitative Research, New Edition*. Princeton University Press, Princeton, 2021.

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A Watermark for Large Language Models, June 2023. URL http://arxiv.org/abs/2301.10226. arXiv:2301.10226 [cs].

J. F. Kiviet. Testing the impossible: Identifying exclusion restrictions. *Journal of Econometrics*, 218(2):294–316, Oct. 2020. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.04.018. URL https://www.sciencedirect.com/science/article/pii/S030440762030138X.

G. Klein, M. Jalaeian, R. R. Hoffman, and S. T. Mueller. The plausibility transition model for sensemaking. *Frontiers in Psychology*, 14, 2023. ISSN 1664-1078. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1160132.

J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2018340118. URL https://pnas.org/doi/full/10.1073/pnas.2018340118.

W. Knight. The Joy and Dread of AI Image Generators Without Limits. *Wired*, Sept. 2022. ISSN 1059-1028. URL https://www.wired.com/story/the-joy-and-dread-of-ai-image-generators-without-limits/.

B. Koch, E. Denton, A. Hanna, and J. G. Foster. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, Dec. 2021. URL http://arxiv.org/abs/2112.01716. arXiv:2112.01716 [cs, stat].

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis, Mar. 2021. URL http://arxiv.org/abs/2009.09761. arXiv:2009.09761 [cs, eess, stat].

L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice*, 2(2):14:1–14:16, Apr. 2021. doi: 10.1145/3412869. URL https://dl.acm.org/doi/10.1145/3412869.

E. Kosoy, J. Collins, D. Chan, J. Hamrick, S. H. Huang, J. F. Canny, and A. Gopnik. Bridging AI and Cognitive Science (BAICS), 2020. URL https://baicsworkshop.github.io/program/baics_24.html.

S. Kreps, R. M. McCain, and M. Brundage. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9(1):104–117, 2022. ISSN 2052-2630, 2052-2649. doi: 10.1017/XPS.2020.37. URL https://www.cambridge.org/core/product/identifier/S2052263020000378/type/journal_article.

J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. C. Balli, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, Jan. 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00447. URL http://arxiv.org/abs/2103.12028. arXiv:2103.12028 [cs].

R. Krishnan. FraudGPT: The Villain Avatar of ChatGPT, July 2023. URL https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt.

A. Lashbrook. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind, Aug. 2018. URL https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

S. Lazar and A. Nelson. AI safety on whose terms? *Science*, 381(6654):138–138, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi8982. URL https://www.science.org/doi/10.1126/science.adi8982.

J. Lee, M. Hameleers, and S. Y. Shin. The emotional effects of multimodal disinformation: How multimodality, issue relevance, and anxiety affect misperceptions about the flu vaccine. *New Media & Society*, page 146144482311539, Apr. 2023a. ISSN 1461-4448, 1461-7315. doi: 10.1177/14614448231153959. URL http://journals.sagepub.com/doi/10.1177/14614448231153959.

M. Lee, M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X. L. Li, F. Ladhak, F. Rong, R. E. Wang, M. Kwon, J. S. Park, H. Cao, T. Lee, R. Bommasani, M. Bernstein, and P. Liang. Evaluating Human-Language Model Interaction, Sept. 2023b. URL http://arxiv.org/abs/2212.09746. arXiv:2212.09746 [cs].

N. Lee, W. Ping, P. Xu, M. Patwary, P. Fung, M. Shoeybi, and B. Catanzaro. Factuality Enhanced Language Models for Open-Ended Text Generation, Mar. 2023c. URL http://arxiv.org/abs/2206.04624. arXiv:2206.04624 [cs].

S. Lee, H. Gil De Zúñiga, and K. Munger. Antecedents and consequences of fake news exposure: a two-panel study on how news use and different indicators of fake news exposure affect media trust. *Human Communication Research*, 49(4):408–420, Sept. 2023d. ISSN 0360-3989, 1468-2958. doi: 10.1093/hcr/hqad019. URL https://academic.oup.com/hcr/article/49/4/408/7111256.

N. G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Jan. 2012. ISBN 9780262298247. doi: 10.7551/mitpress/8179.001.0001. URL https://direct.mit.edu/books/oa-monograph/2908/Engineering-a-Safer-WorldSystems-Thinking-Applied.

J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models, May 2023a. URL http://arxiv.org/abs/2305.11747. arXiv:2305.11747 [cs].

Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating Object Hallucination in Large Vision-Language Models, May 2023b. URL http://arxiv.org/abs/2305.10355. arXiv:2305.10355 [cs].

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic Evaluation of Language Models, Nov. 2022. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs] version: 1.

W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. GPT detectors are biased against non-native English writers, July 2023. URL http://arxiv.org/abs/2304.02819. arXiv:2304.02819 [cs].

Q. V. Liao and J. W. Vaughan. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap, Aug. 2023. URL http://arxiv.org/abs/2306.01941. arXiv:2306.01941 [cs].

Q. V. Liao and Z. Xiao. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap, June 2023. URL http://arxiv.org/abs/2306.03100. arXiv:2306.03100 [cs].

T. Liao, R. Taori, D. Raji, and L. Schmidt. 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*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/757b505cfd34c64c85ca5b5690ee5293-Paper-round2.pdf.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 9783319106021. doi: 10.1007/978-3-319-10602-1_48.

J. Lise, S. Seitz, and J. Smith. Equilibrium Policy Experiments and the Evaluation of Social Programs, Feb. 2004. URL https://www.nber.org/papers/w10283.

C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Jan. 2017. URL http://arxiv.org/abs/1603.08023. arXiv:1603.08023 [cs].

N. F. Liu, T. Zhang, and P. Liang. Evaluating Verifiability in Generative Search Engines, Apr. 2023a. URL http://arxiv.org/abs/2304.09848. arXiv:2304.09848 [cs].

T. Liu, Y. Zhang, C. Brockett, Y. Mao, Z. Sui, W. Chen, and B. Dolan. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation, Apr. 2022. URL http://arxiv.org/abs/2104.08704. arXiv:2104.08704 [cs].

Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment, Aug. 2023b. URL http://arxiv.org/abs/2308.05374. arXiv:2308.05374 [cs].

M. Lodge and C. S. Taber. *The Rationalizing Voter*. Cambridge University Press, Apr. 2013. ISBN 9780521763509. Google-Books-ID: tqI5GxWmIU4C.

P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7 (1):74–101, Jan. 2023. ISSN 2397-3374. doi: 10.1038/s41562-022-01460-1. URL https://www.nature.com/articles/s41562-022-01460-1.

J. Lovato, L. Hébert-Dufresne, J. St-Onge, R. Harp, G. S. Lopez, S. P. Rogers, I. U. Haq, and J. Onaolapo. Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks, Jan. 2023. URL http://arxiv.org/abs/2210.10026. arXiv:2210.10026 [physics].

C. Lu, J. Kay, and K. McKee. Subverting machines, fluctuating identities: Re-learning human categorization. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1005–1015, Seoul Republic of Korea, June 2022. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533161. URL https://dl.acm.org/doi/10.1145/3531146.3533161.

Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images, Sept. 2023. URL http://arxiv.org/abs/2304.13023. arXiv:2304.13023 [cs].

A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, Nov. 2022. URL https://ui.adsabs.harvard.edu/abs/2022arXiv221102001S. ADS Bibcode: 2022arXiv221102001S.

A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable Bias: Analyzing Societal Representations in Diffusion Models, Mar. 2023. URL http://arxiv.org/abs/2303.11408. arXiv:2303.11408 [cs].

X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems. *Pattern Recognition*, 110:107332, Feb. 2021. ISSN 00313203. doi: 10.1016/j.patcog.2020.107332. URL http://arxiv.org/abs/1907.10456. arXiv:1907.10456 [cs, eess].

A. Mandal, S. Leavy, and S. Little. Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models, Apr. 2023. URL http://arxiv.org/abs/2304.13855. arXiv:2304.13855 [cs].

M. Marchiori Manerba, R. Guidotti, L. Passaro, and S. Ruggieri. Bias Discovery within Human Raters: A Case Study of the Jigsaw Dataset. In *Proceedings of the LREC 2022 workshop on Perspectivist Approaches to Disagreement in NLP (NLPerspectives)*, Paris, June 2022. European Language Resources Association (ELRA). URL https://aclanthology.org/2022.nlperspectives-1.4.pdf.

V. Marda and S. Narayan. Data in New Delhi's predictive policing system. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 317–324, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372865. URL https://doi.org/10.1145/3351095.3372865.

C. Martel, G. Pennycook, and D. G. Rand. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):47, Oct. 2020. ISSN 2365-7464. doi: 10.1186/s41235-020-00252-3. URL https://doi.org/10.1186/s41235-020-00252-3.

A. Marwick and R. Lewis. *Media Manipulation and Disinformation Online*. Data & Society, 2017. URL https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf.

J. N. Matias. Humans and algorithms work together — so study them together. *Nature*, 617(7960): 248–251, May 2023. doi: 10.1038/d41586-023-01521-z. URL https://www.nature.com/articles/d41586-023-01521-z.

P. Mattson, A. Selvan, D. Kanter, V. Janapa Reddi, R. Roberts, and J. Corbo. Perspective: Unlocking ML requires an ecosystem approach, Mar. 2023. URL https://mlcommons.org/.

S. Matz, J. Teeny, S. S. Vaid, G. M. Harari, and M. Cerf. The Potential of Generative AI for Personalized Persuasion at Scale. preprint, PsyArXiv, Apr. 2023. URL https://osf.io/rn97c.

J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On Faithfulness and Factuality in Abstractive Summarization, May 2020. URL http://arxiv.org/abs/2005.00661. arXiv:2005.00661 [cs].

MediaWell. "Please do not include us": Workshop on AI Ethics and Inclusion – MediaWell, 2019. URL https://mediawell.ssrc.org/?post_type=tribe_events&p=52376.

J. Metcalf, E. Moss, and D. Boyd. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, 86(2):449–476, Aug. 2019. ISSN 1944-768X. doi: 10.1353/sor.2019.0022. URL https://muse.jhu.edu/pub/1/article/732185.

R. Metz. OpenAI Shut Down ChatGPT to Fix Bug Exposing User Chat Titles. *Bloomberg.com*, Mar. 2023. URL https://www.bloomberg.com/news/articles/2023-03-21/openai-shut-down-chatgpt-to-fix-bug-exposing-user-chat-titles.

A. Milanez. The impact of AI on the workplace: Evidence from OECD case studies of AI implementation. Technical report, Organisation for Economic Co-operation and Development, Paris, Mar. 2023. URL https://www.oecd-ilibrary.org/social-issues-migration-health/the-impact-of-ai-on-the-workplace-evidence-from-oecd-case-studies-of-ai-implementation_2247ce58-en.

R. Millière. Adversarial Attacks on Image Generation With Made-Up Words, Aug. 2022. URL http://arxiv.org/abs/2208.04135. arXiv:2208.04135 [cs].

S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, May 2023. URL http://arxiv.org/abs/2305.14251. arXiv:2305.14251 [cs].

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

S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, and C. Ahuja. FACTIFY: A Multi-Modal Fact Verification Dataset. Vancouver, Canada, 2022.

M. Mitchell. How do we know how smart AI systems are? *Science*, 381(6654):adj5957, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adj5957. URL https://www.science.org/doi/10.1126/science.adj5957.

S. Mohamed, M.-T. Png, and W. Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, Dec. 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00405-8. URL https://doi.org/10.1007/s13347-020-00405-8.

M. Mozes, J. Hoffmann, K. Tomanek, M. Kouate, N. Thain, A. Yuan, T. Bolukbasi, and L. Dixon. Towards Agile Text Classifiers for Everyone, Feb. 2023. URL http://arxiv.org/abs/2302.06541. arXiv:2302.06541 [cs].

D. Muhlgay, O. Ram, I. Magar, Y. Levine, N. Ratner, Y. Belinkov, O. Abend, K. Leyton-Brown, A. Shashua, and Y. Shoham. Generating Benchmarks for Factuality Evaluation of Language Models, July 2023. URL http://arxiv.org/abs/2307.06908. arXiv:2307.06908 [cs].

J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, May 2023. ISSN 2730-5961. doi: 10.1007/s43681-023-00289-2. URL https://doi.org/10.1007/s43681-023-00289-2.

R. Naik and B. Nushi. Social Biases through the Text-to-Image Generation Lens, Mar. 2023. URL http://arxiv.org/abs/2304.06034. arXiv:2304.06034 [cs].

K. Nakamura, S. Levy, and W. Y. Wang. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France, May 2020. European Language Resources Association. ISBN 9791095546344. URL https://aclanthology.org/2020.lrec-1.755.

A. Narayanan and S. Kapoor. GPT-4 and professional benchmarks: the wrong answer to the wrong question, 2023. URL https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

National Institute of Standards and Technology. U.S. Leadership in AI. Technical report, 2019. URL https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_federngagement_plan_9aug2019.pdf.

National Institute of Standards and Technology. AI RMF Playbook. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, 2021a. URL https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

National Institute of Standards and Technology. AI Risk Management Framework. *NIST*, July 2021b. URL https://www.nist.gov/itl/ai-risk-management-framework.

M. K. Ngueajio and G. Washington. Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. volume 13518, pages 421–440. 2022. URL http://arxiv.org/abs/2211.09511. arXiv:2211.09511 [cs, eess].

S. J. Nightingale and H. Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, Feb. 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2120481119. URL https://pnas.org/doi/full/10.1073/pnas.2120481119.

H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of GPT-4 on Medical Challenge Problems, Apr. 2023. URL http://arxiv.org/abs/2303.13375. arXiv:2303.13375 [cs].

J. Novikova, O. Dušek, and V. Rieser. RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, 2018. doi: 10.18653/v1/N18-2012. URL http://arxiv.org/abs/1803.05928. arXiv:1803.05928 [cs].

S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh2586. URL https://www.science.org/doi/10.1126/science.adh2586.

A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, Sept. 2016. URL http://arxiv.org/abs/1609.03499. arXiv:1609.03499 [cs] version: 2.

OpenAI. GPT-4 System Card. Technical report, Mar. 2023a. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf.

OpenAI. GPT-4 Technical Report, Mar. 2023b. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

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

Organisation for Economic Co-operation and Development. Algorithmic Impact Assessment tool - OECD.AI, a. URL https://oecd.ai/en/catalogue/tools/algorithmic-impact-assessment-tool.

Organisation for Economic Co-operation and Development. Expert Group on AI Incidents - OECD.AI, b. URL https://oecd.ai/en//network-of-experts/working-group/10836.

Organisation for Economic Co-operation and Development. Artificial intelligence & responsible business conduct. Technical report, Organisation for Economic Co-operation and Development, Paris, c. URL https://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf.

P. Organisciak, M. Efron, K. Fenlon, and M. Senseney. Evaluating rater quality and rating difficulty in online annotation activities: Evaluating Rater Quality and Rating Difficulty in Online Annotation Activities. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012. ISSN 00447870. doi: 10.1002/meet.14504901166. URL https://onlinelibrary.wiley.com/doi/10.1002/meet.14504901166.

M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation, Apr. 2023. URL http://arxiv.org/abs/2304.01816. arXiv:2304.01816 [cs].

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

R. Owen, M. Pansera, P. Macnaghten, and S. Randles. Organisational institutionalisation of responsible innovation. *Research Policy*, 50(1):104132, Jan. 2021. ISSN 0048-7333. doi: 10.1016/j.respol.2020.104132. URL https://www.sciencedirect.com/science/article/pii/S0048733320302079.

S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis. Synthetic Misinformers: Generating and Combating Multimodal Misinformation, Mar. 2023. URL http://arxiv.org/abs/2303.01217. arXiv:2303.01217 [cs].

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, Aug. 2023a. URL http://arxiv.org/abs/2304.03442. arXiv:2304.03442 [cs].

P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI Deception: A Survey of Examples, Risks, and Potential Solutions, Aug. 2023b. URL http://arxiv.org/abs/2308.14752. arXiv:2308.14752 [cs].

S. Park, C. Fisher, T. Flew, and U. Dulleck. Global Mistrust in News: The Impact of Social Media on Trust. *International Journal on Media Management*, 22(2):83–96, Apr. 2020. ISSN 1424-1277, 1424-1250. doi: 10.1080/14241277.2020.1799794. URL https://www.tandfonline.com/doi/full/10.1080/14241277.2020.1799794.

F. Pasquale and G. Malgieri. Opinion | If You Don't Trust A.I. Yet, You're Not Wrong. *The New York Times*, July 2021. ISSN 0362-4331. URL https://www.nytimes.com/2021/07/30/opinion/artificial-intelligence-european-union.html.

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon Emissions and Large Neural Network Training, Apr. 2021. URL http://arxiv.org/abs/2104.10350. arXiv:2104.10350 [cs].

M. Pawelec. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society*, 1(2): 19, Sept. 2022. ISSN 2731-4669. doi: 10.1007/s44206-022-00010-6. URL https://doi.org/10.1007/s44206-022-00010-6.

S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot, Feb. 2023. URL http://arxiv.org/abs/2302.06590. arXiv:2302.06590 [cs].

G. Pennycook, T. D. Cannon, and D. G. Rand. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880, Dec. 2018. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0000465. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000465.

G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, 66(11):4944–4957, Nov. 2020. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.2019.3478. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2019.3478.

I. Pentina, T. Hancock, and T. Xie. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600, Mar. 2023. ISSN 0747-5632. doi: 10.1016/j.chb.2022.107600. URL https://www.sciencedirect.com/science/article/pii/S0747563222004204.

E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red Teaming Language Models with Language Models, Feb. 2022a. URL http://arxiv.org/abs/2202.03286. arXiv:2202.03286 [cs].

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

Y. Perlitz, E. Bandel, A. Gera, O. Arviv, L. Ein-Dor, E. Shnarch, N. Slonim, M. Shmueli-Scheuer, and L. Choshen. Efficient Benchmarking (of Language Models), Sept. 2023. URL http://arxiv.org/abs/2308.11696. arXiv:2308.11696 [cs].

B. Perrigo. Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer, Jan. 2023. URL https://time.com/6247678/openai-chatgpt-kenya-workers/.

Perspective API. Perspective API. URL https://perspectiveapi.com/.

R. Pfefferkorn. 'Deepfakes' in the Courtroom, Oct. 2020. URL https://papers.ssrn.com/abstract=4321140.

V. Prabhakaran, A. M. Davani, and M. Díaz. On Releasing Annotator-Level Labels and Information in Datasets, Oct. 2021. URL http://arxiv.org/abs/2110.05699. arXiv:2110.05699 [cs].

C. Prunkl and J. Whittlestone. Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society, Jan. 2020. URL http://arxiv.org/abs/2001.04335. arXiv:2001.04335 [cs].

R. Qadri, R. Shelby, C. L. Bennett, and E. Denton. 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*, pages 506–517, Chicago IL USA, June 2023a. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594016. URL https://dl.acm.org/doi/10.1145/3593013.3594016.

R. Qadri, R. Shelby, and E. Denton. Towards Globally Responsible and Human-Centered Text-to-Image Evaluations. In *Practical ML for Developing Countries Workshop @ ICLR 2023*, 2023b. URL https://pml4dc.github.io/iclr2023/pdf/PML4DC_ICLR2023_32.pdf.

Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models, Aug. 2023. URL http://arxiv.org/abs/2305.13873. arXiv:2305.13873 [cs].

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. v. d. Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev,

D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. d. M. d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. d. L. Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling Language Models: Methods, Analysis & Insights from Training Gopher, Jan. 2022. URL http://arxiv.org/abs/2112.11446. arXiv:2112.11446 [cs].

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. 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, pages 33–44, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna. AI and the Everything in the Whole Wide World Benchmark, Nov. 2021. URL http://arxiv.org/abs/2111.15366. arXiv:2111.15366 [cs].

I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 959–972, New York, NY, USA, June 2022a. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533158. URL https://dl.acm.org/doi/10.1145/3531146.3533158.

I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, June 2022b. URL http://arxiv.org/abs/2206.04737. arXiv:2206.04737 [cs].

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation, Feb. 2021. URL http://arxiv.org/abs/2102.12092. arXiv:2102.12092 [cs].

K. Ramesh, A. R. KhudaBukhsh, and S. Kumar. 'Beach' to 'Bitch': Inadvertent Unsafe Transcription of Kids' Content on YouTube. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (11):12108–12118, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i11.21470. URL https://ojs.aaai.org/index.php/AAAI/article/view/21470.

J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr. Red-Teaming the Stable Diffusion Safety Filter, Nov. 2022. URL http://arxiv.org/abs/2210.04610. arXiv:2210.04610 [cs].

H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. Measuring Attribution in Natural Language Generation Models, Aug. 2022. URL http://arxiv.org/abs/2112.12870. arXiv:2112.12870 [cs].

M. Rauh, J. Mellor, J. Uesato, P.-S. Huang, J. Welbl, L. Weidinger, S. Dathathri, A. Glaese, G. Irving, I. Gabriel, W. Isaac, and L. A. Hendricks. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. *Advances in Neural Information Processing Systems*, 35: 24720–24739, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9ca22870ae0ba55ee50ce3e2d269e5de-Abstract-Datasets_and_Benchmarks.html.

R. Rini. Weaponized Skepticism: An Analysis of Social Media Deception as Applied Political Epistemology. In *Political Epistemology*, pages 31–48. Oxford University Press, May 2021. ISBN 9780192893338 9780191914607. doi: 10.1093/oso/9780192893338.003.0003. URL https://academic.oup.com/book/39301/chapter/338889437.

S. Rismani, R. Shelby, A. Smart, R. Delos Santos, A. Moon, and N. Rostamzadeh. Beyond the ml model: Applying safety engineering frameworks to text-to-image development. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 70–83, 2023.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object Hallucination in Image Captioning, Mar. 2019. URL http://arxiv.org/abs/1809.02156. arXiv:1809.02156 [cs].

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, Apr. 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

A. Roy and R. Umbach. Measuring User Journey Friction in Search Engines. *Defensive Publications Series*, Mar. 2023. URL https://www.tdcommons.org/dpubs_series/5761.

A. Roy, E. Jin, and R. Umbach. Determining User Journey Risk Trajectories in Information Seeking Sessions. *Defensive Publications Series*, July 2023. URL https://www.tdcommons.org/dpubs_series/6082.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. 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, June 2018. Association for Computational Linguistics.

J. Rudolph, S. Tan, and S. Tan. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1):342–363, Jan. 2023. ISSN 2591-801X. doi: 10.37074/jalt.2023.6.1.9. URL https://journals.sfu.ca/jalt/index.php/jalt/article/view/689.

G. Russo, N. Stoehr, and M. H. Ribeiro. ACTI at EVALITA 2023: Overview of the Conspiracy Theory Identification Task, Sept. 2023. URL http://arxiv.org/abs/2307.06954. arXiv:2307.06954 [cs].

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, May 2022. URL http://arxiv.org/abs/2205.11487. arXiv:2205.11487 [cs].

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. June 2016. URL http://arxiv.org/abs/1606.03498. arXiv:1606.03498 [cs].

N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada, Mar. 2021. ACM. ISBN 9781450383097. doi: 10.1145/3442188.3445896. URL https://dl.acm.org/doi/10.1145/3442188.3445896.

M. Sap, R. LeBras, D. Fried, and Y. Choi. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, Oct. 2022a. URL https://arxiv.org/abs/2210.13312v2.

M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection, May 2022b. URL http://arxiv.org/abs/2111.07997. arXiv:2111.07997 [cs].

A. Satariano and P. Mozur. The People Onscreen Are Fake. The Disinformation Is Real. *The New York Times*, Feb. 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.

R. Schaeffer, B. Miranda, and S. Koyejo. Are Emergent Abilities of Large Language Models a Mirage?, May 2023. URL http://arxiv.org/abs/2304.15004. arXiv:2304.15004 [cs].

M. K. Scheuerman, A. Hanna, and E. Denton. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2):317:1–317:37, Oct. 2021. doi: 10.1145/3476058. URL https://dl.acm.org/doi/10.1145/3476058.

D. Schlangen. Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research, Aug. 2019. URL http://arxiv.org/abs/1908.10747. arXiv:1908.10747 [cs].

A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA, Jan. 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

M. P. Sendak, W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, M. Nichols, M. Revoir, F. Yashar, C. Miller, K. Kester, S. Sandhu, K. Corey, N. Brajer, C. Tan, A. Lin, T. Brown, S. Engelbosch, K. Anstrom, M. C. Elish, K. Heller, R. Donohoe, J. Theiling, E. Poon, S. Balu, A. Bedoya, and C. O'Brien. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics*, 8(7):e15182, July 2020. doi: 10.2196/15182. URL https://medinform.jmir.org/2020/7/e15182.

Z. Sha, Z. Li, N. Yu, and Y. Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models, Jan. 2023. URL http://arxiv.org/abs/2210.06998. arXiv:2210.06998 [cs].

S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017.

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

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

R. Shiffrin and M. Mitchell. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120, Mar. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2300963120. URL https://pnas.org/doi/10.1073/pnas.2300963120.

M. Shin, J. Kim, and M. Kim. Human Learning from Artificial Intelligence: Evidence from Human Go Players' Decisions after AlphaGo. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43), 2021. URL https://escholarship.org/uc/item/6q05n7pz.

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

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

F. Soldner, V. Pérez-Rosas, and R. Mihalcea. Box of Lies: Multimodal Deception Detection in Dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1175. URL https://aclanthology.org/N19-1175.

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

G. Spitale, N. Biller-Andorno, and F. Germani. AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850, June 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adh1850. URL https://www.science.org/doi/10.1126/sciadv.adh1850.

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, (...), and Z. Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2023. URL http://arxiv.org/abs/2206.04615. arXiv:2206.04615 [cs, stat].

R. Steed and A. Caliskan. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 701–713, Mar. 2021. doi: 10.1145/3442188.3445932. URL http://arxiv.org/abs/2010.15052. arXiv:2010.15052 [cs].

J. Stilgoe, R. Owen, and P. Macnaghten. Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580, Nov. 2013. ISSN 0048-7333. doi: 10.1016/j.respol.2013.05.008. URL https://www.sciencedirect.com/science/article/pii/S0048733313000930.

T. Stoev, K. Yordanova, and E. L. Tonkin. 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)*, pages 534–539, Atlanta, GA, USA, Mar. 2023. IEEE. ISBN 9781665453813. doi: 10.1109/PerComWorkshops56833.2023.10150364. URL https://ieeexplore.ieee.org/document/10150364/.

S. S. Sundar, M. D. Molina, and E. Cho. Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? *Journal of Computer-Mediated Communication*, 26(6):301–319, Nov. 2021. ISSN 1083-6101. doi: 10.1093/jcmc/zmab010. URL https://academic.oup.com/jcmc/article/26/6/301/6336055.

H. Suresh and J. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO

'21, pages 1–9, New York, NY, USA, Nov. 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483305. URL https://dl.acm.org/doi/10.1145/3465416.3483305.

E. Svikhnushina and P. Pu. Approximating Online Human Evaluation of Social Chatbots with Prompting, Aug. 2023. URL http://arxiv.org/abs/2304.05253. arXiv:2304.05253 [cs].

M. Tahaei, M. Constantinides, D. Quercia, S. Kennedy, M. Muller, S. Stumpf, Q. V. Liao, R. Baeza-Yates, L. Aroyo, J. Holbrook, E. Luger, M. Madaio, I. G. Blumenfeld, M. De-Arteaga, J. Vitak, and A. Olteanu. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, pages 1–4, New York, NY, USA, Apr. 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3583178. URL https://doi.org/10.1145/3544549.3583178.

M. B. Tannenbaum, J. Hepler, R. S. Zimmerman, L. Saul, S. Jacobs, K. Wilson, and D. Albarracín. Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, 141(6):1178–1204, Nov. 2015. ISSN 1939-1455, 0033-2909. doi: 10.1037/a0039729. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0039729.

A. H. Taylor, A. P. Bastos, R. L. Brown, and C. Allen. The signature-testing approach to mapping biological and artificial intelligences. *Trends in Cognitive Sciences*, 26(9):738–750, Sept. 2022. ISSN 13646613. doi: 10.1016/j.tics.2022.06.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661322001334.

L. Taylor. Amnesty International criticised for using AI-generated images. *The Guardian*, May 2023. ISSN 0261-3077. URL https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism.

The White House. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, July 2023. URL https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. LaMDA: Language Models for Dialog Applications, Feb. 2022. URL http://arxiv.org/abs/2201.08239. arXiv:2201.08239 [cs].

S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, and E. Gómez. Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, June 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12647. URL https://www.jair.org/index.php/jair/article/view/12647.

N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed. 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, pages 254–265, New York, NY, USA, July 2021. Association

for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462540. URL https://dl.acm.org/doi/10.1145/3461702.3462540.

C. Toups, R. Bommasani, K. A. Creel, S. H. Bana, D. Jurafsky, and P. Liang. Ecosystem-level Analysis of Deployed Machine Learning Reveals Homogeneous Outcomes, July 2023. URL http://arxiv.org/abs/2307.05862. arXiv:2307.05862 [cs].

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

S. Tu, C. Li, J. Yu, X. Wang, L. Hou, and J. Li. ChatLog: Recording and Analyzing ChatGPT Across Time, Apr. 2023. URL http://arxiv.org/abs/2304.14106. arXiv:2304.14106 [cs].

S. Turkle. *Alone together: Why we expect more from technology and less from each other.* Basic Books, New York, NY, US, 2011.

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

T. Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, Feb. 2023. URL https://arxiv.org/abs/2302.08399v5.

US Chief Information Officers Council. Algorithmic Impact Assessment. URL https://www.cio.gov/aia-eia-js/#/.

C. Vaccari and A. Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1): 205630512090340, Jan. 2020. ISSN 2056-3051, 2056-3051. doi: 10.1177/2056305120903408. URL http://journals.sagepub.com/doi/10.1177/2056305120903408.

E. van Miltenburg, W.-T. Lu, E. Krahmer, A. Gatt, G. Chen, L. Li, and K. van Deemter. Gradations of Error Severity in Automatic Image Descriptions. In *Proceedings of The 13th International Conference on Natural Language Generation*, pages 398–411, 2020.

H. Vasconcelos, G. Bansal, A. Fourney, Q. V. Liao, and J. W. Vaughan. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions, Feb. 2023. URL http://arxiv.org/abs/2302.07248. arXiv:2302.07248 [cs].

J. W. Vaughan and H. Wallach. A Human-Centered Agenda for Intelligible Machine Learning. *In Machines We Trust: Perspectives on Dependable AI*, Dec. 2021. URL https://www.microsoft.com/en-us/research/publication/a-human-centered-agenda-for-intelligible-machine-learning/.

P. Verma. They thought loved ones were calling for help. It was an AI scam. *Washington Post*, Mar. 2023. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.

V. Veselovsky, M. H. Ribeiro, and R. West. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks, June 2023. URL http://arxiv.org/abs/2306.07899. arXiv:2306.07899 [cs].

J. Vincent. Republicans respond to Biden reelection announcement with AI-generated attack ad, Apr. 2023. URL https://www.theverge.com/2023/4/25/23697328/biden-reelection-rnc-ai-generated-attack-ad-deepfake.

M. Vogel and J. Manyika. National artificial intelligence advisory committee report, year 1. Working Paper, National Artificial Intelligence Advisory Committee, May 2023. URL https://ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf.

C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03666-1. URL https://www.nature.com/articles/s41586-021-03666-1.

D. Wan and M. Bansal. Evaluating and Improving Factuality in Multimodal Abstractive Summarization, Nov. 2022. URL http://arxiv.org/abs/2211.02580. arXiv:2211.02580 [cs].

A. Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, and O. Russakovsky. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets, July 2021. URL http://arxiv.org/abs/2004.07999. arXiv:2004.07999 [cs].

A. Wang, S. Barocas, K. Laird, and H. Wallach. Measuring Representational Harms in Image Captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 324–335, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533099. URL https://dl.acm.org/doi/10.1145/3531146.3533099.

S. Wang, N. Cooper, M. Eby, and E. S. Jo. From Human-Centered to Social-Centered Artificial Intelligence: Assessing ChatGPT's Impact through Disruptive Events, May 2023. URL http://arxiv.org/abs/2306.00227. arXiv:2306.00227 [cs].

Y. Wang, Q. Wang, S. Shi, X. He, Z. Tang, K. Zhao, and X. Chu. Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 744–751, Melbourne, Australia, May 2020. IEEE. ISBN 9781728160955. doi: 10.1109/CCGrid49817.2020.00-15. URL https://ieeexplore.ieee.org/document/9139681/.

C. Wardle and H. Derakhshan. Information Disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report DGI(2017)09, 2017. URL https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf.

M. Webb. The Impact of Artificial Intelligence on the Labor Market, Nov. 2019. URL https://papers.ssrn.com/abstract=3482150.

A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How Does LLM Safety Training Fail?, July 2023. URL http://arxiv.org/abs/2307.02483. arXiv:2307.02483 [cs].

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent Abilities of Large Language Models, Oct. 2022. URL http://arxiv.org/abs/2206.07682. arXiv:2206.07682 [cs].

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

T. Weikmann and S. Lecheler. Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, page 146144482211416, Dec. 2022. ISSN 1461-4448, 1461-7315. doi: 10.1177/14614448221141648. URL http://journals.sagepub.com/doi/10.1177/14614448221141648.

K. Weise and C. Metz. When A.I. Chatbots Hallucinate. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang. Challenges in Detoxifying Language Models, Sept. 2021. URL http://arxiv.org/abs/2109.07445. arXiv:2109.07445 [cs].

E. Wenger, R. Bhattacharjee, A. N. Bhagoji, J. Passananti, E. Andere, H. Zheng, and B. Zhao. Finding Naturally Occurring Physical Backdoors in Image Datasets. *Advances in Neural Information Processing Systems*, 35:22103–22116, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8af749935131cc8ea5dae4f6d8cdb304-Abstract-Datasets_and_Benchmarks.html.

O. Wiles, I. Albuquerque, and S. Gowal. Discovering Bugs in Vision Models using Off-the-shelf Image Generation and Captioning, May 2023. URL http://arxiv.org/abs/2208.08831. arXiv:2208.08831 [cs, stat].

D. Wilson and T. Wharton. Relevance and prosody. *Journal of Pragmatics*, 38(10):1559–1579, Oct. 2006. ISSN 0378-2166. doi: 10.1016/j.pragma.2005.04.012. URL https://www.sciencedirect.com/science/article/pii/S0378216606000397.

J. Wolff. Fairness, Respect and the Egalitarian "Ethos" Revisited. *The Journal of Ethics*, 14(3/4): 335–350, 2010. ISSN 1382-4554. URL https://www.jstor.org/stable/40928983.

C. S. Wu and U. Bhandary. Detection of Hate Speech in Videos Using Machine Learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 585–590, Dec. 2020. doi: 10.1109/CSCI51800.2020.00104. URL https://ieeexplore.ieee.org/abstract/document/9458005.

X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference, Aug. 2023. URL http://arxiv.org/abs/2303.14420. arXiv:2303.14420 [cs].

C. Xiang. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says, Mar. 2023. URL https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

Y. Xiao and W. Y. Wang. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.236. URL https://aclanthology.org/2021.eacl-main.236.

T. Xie and I. Pentina. *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika*. Jan. 2022. ISBN 9780998133157. URL http://hdl.handle.net/10125/79590.

Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao. SneakyPrompt: Evaluating Robustness of Text-to-image Generative Models' Safety Filters, May 2023. URL http://arxiv.org/abs/2305.12082. arXiv:2305.12082 [cs].

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

L. Yu and V. Rieser. Adversarial Textual Robustness on Visual Dialog. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3422–3438, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.212. URL https://aclanthology.org/2023.findings-acl.212.

Z. Zeng, L. Sha, Y. Li, K. Yang, D. Gašević, and G. Chen. Towards Automatic Boundary Detection for Human-AI Collaborative Hybrid Essay in Education, Aug. 2023. URL http://arxiv.org/abs/2307.12267. arXiv:2307.12267 [cs].

B. Zevenbergen. *Internet users as vulnerable and at-risk human subjects: reviewing research ethics Law for technical internet research*. http://purl.org/dc/dcmitype/Text, University of Oxford, 2020. URL https://ora.ox.ac.uk/objects/uuid:5363f3f8-6c13-4e56-b065-2980bdcce46b.

J. Zhang, S. I. Levitan, and J. Hirschberg. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. In *Proc. Interspeech 2020*, pages 359–363, 2020. URL http://www.cs.columbia.edu/speech/PaperFiles/2020/multimodal_deception_interspeech2020.pdf.

L. Zhang, S. Mille, Y. Hou, D. Deutsch, E. Clark, Y. Liu, S. Mahamood, S. Gehrmann, M. Clinciu, K. Chandu, and J. Sedoc. Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization, June 2023. URL http://arxiv.org/abs/2212.10397. arXiv:2212.10397 [cs].

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, Apr. 2018. URL http://arxiv.org/abs/1804.06876. arXiv:1804.06876 [cs].

J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–20, New York, NY, USA, Apr. 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581318. URL https://dl.acm.org/doi/10.1145/3544548.3581318.

V. Zhou. AI is already taking video game illustrators' jobs in China, Apr. 2023. URL https://restofworld.org/2023/ai-image-china-video-game-layoffs/.

A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson. Universal and Transferable Attacks on Aligned Language Models, 2023. URL https://llm-attacks.org/.