Chapter 7
Sociotechnical Implications of Generative Artificial Intelligence for Information Access

Bhaskar Mitra (ORCID), **Henriette Cramer** (ORCID), **and Olya Gurevich**

Abstract Robust access to trustworthy information is a critical need for society with implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies may enable new ways to access information and improve effectiveness of existing information retrieval systems, but we are only starting to understand and grapple with their long-term social implications. In this chapter, we present an overview of some of the systemic consequences and risks of employing generative AI in the context of information access. We also provide recommendations for evaluation and mitigation and discuss challenges for future research.

7.1 Introduction

Robust access to trustworthy information is a critical need for society including implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies such as Large Language Models (LLMs) may enable new ways to access information and improve effectiveness of existing Information Retrieval (IR) systems. More efficient basic task execution with the help of LLMs can also enable people to focus on the more challenging aspects of information retrieval–related tasks and research. However, the long-term social implications of deploying these technologies in the context of information access are not yet well understood. Existing research has focused on how these models may generate biased and harmful content [1, 16, 64, 74, 118, 153, 235] as well as the environmental costs [16, 25, 56, 161, 162, 241] of developing and deploying these models at scale. In the context of information

B. Mitra (✉)
Microsoft Research, Montréal, QC, Canada
e-mail: bmitra@microsoft.com

H. Cramer · O. Gurevich
PaperMoon AI, San Francisco, CA, USA

access, Shah and Bender [184] have argued that certain framings of LLMs as "search engines" lack the necessary theoretical underpinnings and may constitute as a category error.

In this current work, we present a broader perspective on the sociotechnical implications of generative AI for information access. Our perspective is informed by existing literature and aims to provide a summary of known challenges viewed through a systemic lens that we hope will serve as a useful resource for future critical research in this area. We present a summary of these implications next, followed by recommendations for evaluation and mitigation later in this chapter.

7.2   Implications of Generative AI for Information Access

We present a reflection on the potential sociotechnical implications of generative AI, with an emphasis on LLMs, for information access. Generative AI is still an emerging technology, and our understanding of its sociotechnical impact today, and how it may evolve over time, is fairly limited. Our treatment of this topic is therefore necessarily both incomplete and speculative. We are informed by several recent works [16, 197, 233, 234] that attempt to map the landscape of risks and harms from LLMs. What distinguishes our treatment of this topic relative to this previous literature is the specific focus on information access. There has also been work on the considerations for specific applications of LLMs in IR, such as for generating direct responses to users' expressed information needs [184], which is relevant to our current discussion. However, a thorough exploration of every potential application of LLMs in IR systems is beyond the scope of our current work. Instead, we explore the implications for information access through a broader lens that encompasses considerations for content creation, content retrieval, sociopolitical power dynamics, geopolitical inequities, crowd work, ecology, and future of IR research. We reference relevant previous taxonomies and studies throughout this section to both support our claims and to establish meaningful connections in an attempt to present a more complete and consistent view on this topic to the reader.

We adopt the Consequences-Mechanisms-Risks (CMR) framework proposed by Gausen et al. [70] to structure our presentation. Gausen et al. introduce the CMR framework to support designers and developers of AI (and in general any computational) systems to identify and understand: (i) The systemic consequences of developing and deploying the technology under study in the real world (ii) The mechanisms introduced by the said technology responsible for these consequences (iii) The corresponding risks to relevant stakeholders The framework intentionally explicates the higher-level consequences to motivate viewing the challenges through a more systemic lens. The mechanisms, in turn, focus on more low-level system behaviors and aspects of the technology development process that contribute to the consequences and risks and therefore represent sites for more actionable mitigation. These consequences and mechanisms are mapped to relevant potential risks. Through literature survey, in this work, we identify the consequences, mechanisms,

Table 7.1 Overview of potential negative consequences for information access from generative AI, the related mechanisms introduced by these AI technologies, and corresponding risks

Consequences	Mechanisms	Risks
Information ecosystem disruption (Sect. 7.2.1.1)	Content pollution (Sect. 7.2.1.1)	Risks to society: democracy, health and well-being, and global inequity (Sect. 7.2.2.1)
	The "Game of telephone" effect (Sect. 7.2.1.1)	
	Search engine manipulation (Sect. 7.2.1.1)	
	Degrading retrieval quality (Sect. 7.2.1.1)	
	Direct model access (Sect. 7.2.1.1)	
	The paradox of reuse (Sect. 7.2.1.1)	
Concentration of power (Sect. 7.2.1.2)	Compute and data moat (Sect. 7.2.1.2)	
	AI persuasion (Sect. 7.2.1.2)	
	AI alignment (Sect. 7.2.1.2)	
Marginalization (Sect. 7.2.1.3)	Appropriation of data labor (Sect. 7.2.1.3)	
	Bias amplification (Sect. 7.2.1.3)	
	AI exploitation and doxing (Sect. 7.2.1.3)	
Innovation decay (Sect. 7.2.1.4)	Industry capture (Sect. 7.2.1.4)	Risks to IR research (Sect. 7.2.2.2)
	Pollution of research artefacts (Sect. 7.2.1.4)	
Ecological impact (Sect. 7.2.1.5)	Resource demand and waste (Sect. 7.2.1.5)	Risks to environment (Sect. 7.2.2.3)
	Persuasive advertising (Sect. 7.2.1.5)	

and risks of generative AI in the context of information access and organize them according to the CMR framework as shown in Table 7.1. While we acknowledge that this list of consequences-mechanisms-risks is incomplete, we hope that it provides a summary of the sociotechnical concerns already identified in existing literature and provokes new questions for critical future research.

7.2.1 Consequences and Mechanisms

In the context of information access, we identify five potential categories of negative consequences of generative AI and corresponding mechanisms, which we discuss next.

7.2.1.1 Consequence: Information Ecosystem Disruption

To reflect on the implications of generative AI on information access, we must consider the information ecosystem as a whole, and not constrain our discussion only to the application of these emerging technologies directly in IR systems. This

ecosystem includes different actors and stakeholders such as information seekers, content producers, IR systems developers, advertisers, and other sociopolitical actors. While the information ecosystem is constantly evolving, generative AI holds the potential to significantly disrupt how each of these actors operate on their own and how they relate to other actors and stakeholders. This potential for disruption spans across how content is produced, consumed, monetized, and used toward specific ends. By no means do we want to imply that these plausible changes are inherently bad, but the scale of potential disruptions across the ecosystem should motivate careful and thoughtful considerations before these technologies are deployed at scale. We discuss next some the underlying mechanisms introduced by generative AI that may contribute to these disruptions. We encourage the reader to view these mechanisms not just in isolation but to also consider how they may interact with each other and how that may impact the ecosystem over time.

Mechanism: Content Pollution Generative AI enables low-cost generation of derivative low-quality content at an unprecedented scale. As a consequence, synthetic AI-generated content is rapidly and very widely appearing on the Web [98]. On Amazon,[1] AI-generated content includes scammy derivatives of existing publications [115, 129, 155] and fake travel guides [119]. On YouTube,[2] AI-generated video creators have targeted children [4, 97, 116]. We are also witnessing a proliferation of news Web sites almost entirely generated by AI [179], which are being surfaced in search results [46] and funded by online ads [27]. Even reputable publishers have reportedly published AI-generated articles under fake AI-generated author profiles [57]. Beyond news, other synthetic content such as AI-generated images is starting to pollute search results [5, 58]. According to another recent study [212], a "shocking" amount of content on the Web today is machine-translated text. The promise of machine translation is that it could make more content accessible to wider audiences. However, it also amplifies the influence of (sometimes questionable-quality) language technology choices. For example, Thompson et al. [212] found that more low-quality content—rather than high-quality content—was machine translated into lower-resource languages, likely with the goal of generating ad revenue. Concerns have also been raised about LLMs potentially serving as "Misinformation Superspreaders" [26, 157] as they make it trivially easy to inundate the Web with "firehoses of falsehoods."[3] Hoel [97] points out that AI pollution of our information ecosystems is a "tragedy of the commons" [92].

Pollution of our information ecosystem at such scale has critical implications for people and society. When authoring a document requires significant time and effort, then quality, style, and comprehensiveness are factors that readers may consider in deciding whether and how much to trust its content. However, when the cost

---

[1] https://www.amazon.com/

[2] https://www.youtube.com/

[3] https://en.wikipedia.org/wiki/Firehose_of_falsehood

of writing an extensive article approaches zero, it becomes significantly harder for the reader to make that decision. They may not be able to distinguish between an article created based on extensive research, fact-checking, and thoughtful writing practices and one generated instantly based on a short user prompt. Furthermore, the increasing adoption of these same AI authoring tools by reputable publishers and content producers may homogenize the language and style of content on the Web, making it even more difficult for readers to distinguish them from low-quality AI-generated content whose sole intent is to attract ad revenue or to mislead. Such Web pollution is also a concern for future AI models that require large Web-scale datasets to train on. Including AI-generated content in the training data for new AI models may have significant negative impact on model performance, what has been referred to as "Model collapse" [137, 189], "Model Autophagy Disorder" [6], and "Habsburg AI."[4]

Mechanism: The "Game of Telephone" Effect LLMs have recently been employed in conversational search interfaces. In systems such as Bing Copilot, the LLM has access to relevant Web search results from which it can draw information to produce appropriate responses for the information needs expressed by a user. In this scenario, the LLM performs a complex summarization task extracting relevant information from the retrieved documents to answer the search query. In doing so, the LLM now inserts itself between the user and the retrieved Web results. This shifts the responsibility of inspecting the information in the documents and assessing their relevance, trustworthiness, and surrounding context from the user to the LLM. Further, factual errors and inconsistencies may arise between what the LLM produces and what is in the retrieved documents. Seeing the model through an anthropomorphic lens, these errors are sometimes referred to as "hallucinations." A more technical view may see this as a noisy translation akin to the children's game of telephone.[5] Such errors, often subtle and hard to spot, may contribute to misinformation and reduce robustness of the information access system. While the LLM-generated responses may cite relevant documents, it is unlikely that users diligently click the provided links and verify the information in the response is indeed supported by said sources. Even if the LLM reproduces exact pieces of text from the source documents without error, taking these out of the context of the document may lead to unexpected negative consequences. Such examples have previously been reported [215] in context of extracted answers that search engines display on the Search Engine Result Pages (SERPs) as response to the user query. These issues may become more prevalent if conversational search interfaces become a popular way to access online information.

In a more radical proposal, Metzler et al. [141] have suggested that LLMs could directly replace retrieval systems and respond directly to the user based on information in their training data. LLMs are trained to produce statistically plausible

---

[4] https://twitter.com/jathansadowski/status/1625245803211272194

[5] https://en.wikipedia.org/wiki/Game_of_telephone

text sequences, and any semblance to an information retrieval system is likely an important mis-categorization of these models that we should be wary of [184]. The game of telephone effect is likely to be more intense when LLMs are expected to produce information from their training data and not just the in-context information in its input.

The interjection of the LLM between the user and the search results may have other long-term effects. These interfaces may disincentivize users from the practice of verifying information sources and make them less skilled over time at discerning online misinformation. If users get accustomed to information being presented neatly summarized and disconnected from original sources, the critical cognitive skills necessary to distinguish between trustworthy and untrustworthy information may atrophy.

Mechanism: Search Engine Manipulation New applications of LLMs to the IR stack have exposed new attack vectors. Prompt injection attacks [83, 131, 132] that try to blur the line between instructions and data have garnered specific interest. In these types of attacks, Web site owners may inject what looks like instructions to the LLM. When such documents are retrieved and included in the input of the LLM as augmentation, the LLM may mistake the injected prompt in the document content and be vulnerable to manipulation.

Recently, LLMs have also found application in relevance labeling for search [211]. It is not well understood yet whether this may make the search engine vulnerable to improper ranking manipulation by Web site owners and search engine optimization experts. For example, one may employ the same, or similar, LLMs to reproduce the labeling scheme externally and then adapt their Web site content and design to achieve undue high predicted relevance against queries to rank higher on SERPs.

Other attack vectors may include using LLMs to create effective content farms at low cost to manipulate the ranking of Web results or even use LLMs to artificially simulate users interacting with the search system to fake clicks and other user behavior signals, such as reformulations, which search engines depend on.

Mechanism: Degrading Retrieval Quality LLM usage can negatively impact search result quality in a number of (indirect) ways. LLMs can contribute to new attack vectors, but more worryingly, in some cases, the negative effect may be a result of the LLM behaving exactly as it is supposed to. For example, one potential consequence of using conversational search interfaces, is that the quality of feedback from user behavior signals on SERPs may significantly degrade. Historically, users of commercial Web search engines have given search systems noisy implicit feedback through clicks and other actions on SERPs. These actions are part of the key secret sauce of any modern search systems.

However, conversational interfaces may discourage direct user clicks on Web results and at best provide much weaker satisfaction signal that may be gleaned from the users' next utterance in the conversation. This over time may negatively impact the underlying retrieval quality. This makes it important to invest in methods that can infer user satisfaction with high certainty from the natural-language conversations.

However, methods for such signal interpretation are not yet at the level necessary to mitigate these impacts.

In conversational search interfaces and other applications, such as Microsoft Copilot for M365 [140, 231], the LLM may conduct the search on the user's behalf. In this process, the LLM generates search queries. If these queries differ from those that are likely to be submitted by users, then the underlying search system needs to optimize itself for both real user queries and LLM-generated queries. This may have consequences that are not yet well understood. Optimizing the search system directly to improve the LLMs natural language responses may also have unforeseen outcomes, especially in light of the fact that what makes for a good result set for retrieval-augmentation is not yet fully understood [51].

Mechanism: Direct Model Access Another important consideration is the implications of open foundation models [109]. While centralized systems have their own negative implications, as discussed in Sect. 7.2.1.2, open-access generative AI models without any access moderation also pose certain challenges. For example, there are many classes of harmful intents that systems should refuse to respond to. This may include search queries seeking information on methods to self-harm or cause harms to others or requests to generate harmful (and sometimes illegal) content such as Child Sex Abuse Material (CSAM) or Non-Consensual Intimate Information (NCII). Publicly accessible LLMs trained on large Web corpora may produce such irresponsible content in the absence of moderation. Even if a model is trained to not respond to certain classes of queries, it is likely that there will be leakage, and the safety alignment may also be compromised if the model is further finetuned [171]. Such leakage may also happen in the context of traditional search systems. However, in the latter case, all queries are typically logged, allowing for post hoc analysis and identification of critical gaps in the moderation system. Unfortunately, no such mitigation is possible once these generative AI models are released into the wild.

Mechanism: The Paradox of Reuse Content producers and information access technologies are critically inter-dependent [139, 225]. Web sites such as Wikipedia,[6] Stack Exchange,[7] and Reddit[8] produce critical content that is surfaced by information access platforms (e.g., Web search engines) and contribute to making these platforms significantly more useful to their users. In return, these platforms have historically sent traffic back to the Web sites that contribute to their increased readership, subscriptions, and monetization. However, when search platforms stop directing traffic back to Web sites—e.g., by instead surfacing relevant content directly on the search result pages (SERPs)—the relationship becomes less symbiotic toward the content producers, a phenomenon Taraborelli [206] termed the "paradox of reuse."

---

[6] https://www.wikipedia.org/

[7] https://stackexchange.com/

[8] https://www.reddit.com/

The application of LLMs as conversational information access interfaces is likely to significantly intensify this problem. For example, LLMs such as ChatGPT[9] and Google Gemini[10] may gobble up large quantities of content from Web sites as part of their training data and later regurgitate the same information without any attribution back to the sources. Even when models summarize information from multiple online sources with attribution, e.g., Bing Copilot,[11] they typically de-emphasize the references and reduce the likelihood of the searcher clicking through to the source Web sites as compared to the classic ten-blue-links interface. There is evidence [52] to suggest that this phenomenon is already happening at scale and is jeopardizing the "grand bargain at the heart of the web" [93].

7.2.1.2 Consequence: Concentration of Power

> We may have democracy, or we may have wealth concentrated in the hands of a few, but we can't have both.   – Louis Brandeis
> As quoted by Lonergan [133]

Technology shapes and is shaped by the sociopolitical power structures within which it exists. The 2024 edition of the World Economic Forum's Global Risks Report [240] lists "technological power concentration" as one of the top global risks for the coming decade and as the biggest upward mover in their annual ranking of global risks compared to the previous year. Deliberation on the social consequences of any technology must therefore include critical consideration of how the technology, and general narratives about the said technology, shifts power and re-architects and codifies structures of hierarchy and control. In this context, the politics and values of those in power to oversee what and how technology is built or regulated, especially when they reinforce hierarchy and authoritarianism (e.g. [72]), (e.g. [59, 72, 120]), become important to consider.

A report [106] from the research institute AI Now[12] similarly asserts "the concentration of economic and political power in the hands of the tech industry—Big Tech in particular" as the core challenge posed by AI. They further note that not just the technologies but the narratives (both the hype and the fear-mongering) around them questionably bolster claims of "foundational" advancements and their unassailable equivalence with scientific progress. These concerns are complemented by discourses within the AI community, such as observations by Birhane et al. [21] that the prominent values expressed and operationalized in top-cited AI papers generally have implications in support of centralization of power. Even if platform owners act accountably to civil society, the concentration of power and control in

---

[9] https://chat.openai.com/

[10] https://gemini.google.com/app

[11] https://www.bing.com/chat

[12] https://ainowinstitute.org/

their hands makes them vulnerable to other actors, such as autocratic governments, and allows that power to be potentially abused for oppressive and harmful intents.

The popularization of generative AI can concentrate that power within large companies, since they emerge as some of the only institutions with the resources to develop and deploy these technologies [111]. The application of these technologies for information access may contribute to further concentration and growing inequities of wealth and power; we discuss three mechanisms in the context of generative AI that may contribute to concentration of power and control.

Mechanism: Compute and Data Moat The development of generative AI is heavily reliant on the availability of large swaths of training data and large-scale computing power for training and deployment. Only a handful of institutions, largely in the private sector, own and control these necessary resources while simultaneously evangelizing AI as crucial geopolitical leverage and critical social infrastructure [106]. Increased access to these models has sometimes been touted as potential paths to mitigation [194, 200], where access may range from being heavily restricted over Application Programming Interfaces (APIs) to "open weight" models [128]. The ability to download models with their learned parameters allows others to further adapt for their own applications and opens the door to more meaningful analysis and audit of these models. However, such "open access" also leads to severe limitations that we should recognize. The availability of the trained models does little to challenge the predominant visions put forth by large technology companies of what AI fundamentally should look like.

One potential direction would be to dismantle the data and compute moat by turning them over from private ownership into public infrastructure for independent researchers and developers and those affiliated with smaller institutions. This also illustrates the importance of existing institutions such as archives, libraries, and universities that have reliable, historical data. The availability of public computer infrastructure would allow a broader set of developers to participate in the reimagination and development of diverse approaches to AI and not merely being forced to be satisfied with critiquing and finetuning artefacts produced by other institutions. However, there is no guarantee that without careful planning and incentives, a proliferation of smaller projects will lead to transformative new or more sustainable results.

Democratizing the control over computational resources provides a mechanism of checks and balances on the future directions of AI systems and may allow for challenges to popular narratives and expectations about generative AI such as exponential growth in model size over time. Infrastructure is however also bound to the particular governing system and local underlying goals and processes. Larger investments in existing research institutes or new alternative companies or non-profits might in certain cases lead to faster results.

Similarly, the research community would benefit from easier access to industry models and APIs for critical studies and auditing. However, access to models or APIs alone is significantly limiting unless that access is also extended to the user-facing systems in which these technologies are deployed. The corresponding

instrumentation data would provide context on how these systems are used by people and potential consequences. This can lead to practical privacy and security questions for platform teams. Practical support for decision-making and, for example, the creation of standards to de-risk those concerns can help alleviate some of those concerns.

Mechanism: AI Persuasion There is an emerging recognition of the dangers of *AI persuasion* [30, 34, 62, 159], which Burtell and Woodside [30] define as "a process by which AI systems alter the beliefs of their users." AI systems may persuade users by appealing to their reason and argument or by using their cognitive biases and heuristics [62]. El-Sayed et al. [62] identify six mechanisms of generative AI persuasion—namely: (i) Trust and rapport (ii) Anthropomorphism (iii) Personalization (iv) Deception and lack of transparency (v) Manipulative strategies (vi) Alteration of choice environment—and corresponding model features that contribute to these mechanisms. In the context of information access and advertising, these capabilities of generative AI can be powerful tools to hyper-target users and steer their behaviors.

Modern online information access and communication platforms monetized with targeted advertising have been said to usher in an age of surveillance capitalism [247, 248]. Information access systems increasingly collect detailed user behavior data that allow them to build accurate user profiles for audience targeting. There is strong evidence that people are more likely to consume information that opposes their own personal views and beliefs when it employs language similar to their own political leanings [243]. So combining users' private preferences and behavioral data with the capabilities of generative AI to produce persuasive language could create worrying tools for mass behavioral manipulation. The impact of such pervasive *algorithmic nudging* [134] may be further pronounced over longer time periods from continuous interactions between the user and the system. Putting these capabilities in the hands of online platform owners, which typically tend to be large multinational for-profit institutions with largely hierarchical non-democratic internal governance structures, poses serious risks to functioning of democratic societies. At the same time, platforms must make decisions about what is acceptable on their platforms to avoid negative user experiences, spam, unwelcoming behavior, and other negative occurrences beyond those outlined in legal compliance alone. Platforms moderate content posted or accessible through the platform [77], and in doing so, they unavoidably impose implementations of values on their users or the values incentivized by, say, advertising needs or other business model–related motivations. For ads, this may mean an incentive to use generative AI to produce hyper-targeted highly personalized persuasive advertisements that convince users to make certain buying decisions. For content, when platforms optimize for increased user engagement, they may knowingly or unknowingly incentivize generative AI models to be producing highly charged content, such as "rage-bait" [101], because it tends to be more persuasive and engaging.

Mechanism: AI Alignment To prevent generative AI models from producing harmful and offensive content, recent research has focused on how to align

model outputs with "human values" [66, 67, 110, 177, 203]. Approaches such as Reinforcement Learning from Human Feedback (RLHF) [38, 246] have been effective in limiting certain types of problematic content from being produced. However, this approach presupposes some notions of desirable values and puts the burden of determining and enforcing them on the shoulders of platform/model developers. Any notions of universal values that might determine what type of content these models should generate—or not generate [214]—are highly contested [20, 105, 167, 170, 180]. Placing these decisions in the exclusive domain of the platform developers, especially in the absence of democratic and civil society oversight, further concentrates power and responsibility. This is not an argument against content moderation itself but against the centralization of control over it without civil oversight or broader societal participation. As a pragmatic example, platforms may not necessarily have the necessary knowledge in-house, making it imperative for them to make successful connections to outside expertise.

7.2.1.3   Consequence: Marginalization

Generative AI, both in its process of development and in its deployment in the context of information access, can marginalize groups and individuals by diminishing their value, power, and well-being. Next, we discuss some of the mechanisms that may contribute to this.

Mechanism: Appropriation of Data Labor Li et al. [123] define *data labor* as "activities that produce digital records useful for capital generation." The term encompasses both witting labor activities—as in the case of crowd work [7], peer production [207, 208], and content moderation [77]—and unwitting activities such as user behavior data and other data generated when users interact with and participate on the platforms. Data labor also encompasses the creation of artefacts by writers [40, 41], artists [220, 221], programmers [219], etc. outside of the AI development process that are nonetheless extracted from the Web and fed in as training data to generative AI models. Appropriation of data labor in this context includes both: (i) The uncompensated appropriation of works by writers, authors, programmers, and peer production communities like Wikipedia [10, 28, 29, 36, 37, 40, 41, 76, 136, 188, 218, 219, 221, 223, 224] (ii) Under-compensated crowd work for data labeling that has been instrumental in the development of these technologies [7, 90, 91, 165, 204, 239, 242]

It is particularly harmful when technology developed on appropriated labor is then employed to displace and automate the jobs of those whose labor was appropriated [8, 47, 223]. Introduction of such automation may involve vicious cycles of perceived skill transfer from people to AI models whereby professional jobs are replaced by corresponding lesser-paid gigified equivalent as auditing and editing of model outputs only [82]. Proprietary AI model capabilities may then continue to improve by learning from workers' inputs, while workers progressively

lose their economic value and power or are even relegated into the role of *moral crumple zones* [63].

This is a critical challenge in the context of information access because: (i) The devaluation of writers and artists have direct implications for the quality of content on the Web (ii) These automated content generation tools are starting to get incorporated directly in information access platforms [166] Similar concerns of commodification and appropriation have also been raised in other information and knowledge access contexts such as in the enterprise [70].

AI for Me, Data Labor for Thee Another pernicious aspect of AI data labor dynamics discussed in the literature is how they can mirror and reify racial capitalism and coloniality, employ global labor exploitation and extractive practices, and reinforce the global north and south divide [19, 45, 88, 114, 149, 154, 202]. While worldwide jobs might be created in certain cases, the workers are typically low paid and deprived of any share of the profit made from technologies built with their labor. These dynamics encompass accruing the benefits of generative AI to privileged populations, while data labor is relegated to already marginalized populations, for example, in the global south. Communities that significantly contribute to AI data labor may even find their own linguistic styles being labeled AI-ese [95] and being forced to repeatedly prove their own humanity [53, 138]. Attempts to bridge the global north-south data gap also in turn may further intensify data extractive practices in the global south [39].

Mechanism: Bias Amplification LLMs and other generative models reproduce and amplify harmful biases and stereotypes from their training datasets [1, 16, 23, 24, 31, 79], which can lead to allocative and representational harms [49]. Harms may also materialize from *demographic blindness* [70] when the model (or the system it is embedded in) treats different individuals and groups as alike when, in fact, it is unwarranted. Examples may include the handling of certain languages as one homogeneous entity without regard for sociolects or dialects [22] or holding different perspectives as equally valid without considerations for historical context or structural dynamics of power. These biases are concerning in the context of information access systems that are responsible for supporting informed citizenry and functioning democracies, health literacy, and knowledge production among other societal needs.

Mechanism: AI Exploitation and Doxing *"AI doxing"* can describe the act of leaking people's private information by an AI system. Weidinger et al. [233] note that this may be caused by models leaking private information (e.g., address and telephone number) present in their training data [33] or when these models are employed to predict people's sensitive attributes (e.g., political and sexual identities) based on what is known about them publicly [117, 158, 172, 244]. Private information in the training data is a challenge even if datasets have been sourced from the public Web because models may continue to regurgitate that information after it has been removed from the Web or bypass safety measures that would prevent such information from surfacing through Web search—e.g., the information

may be protected by robots.txt that blocks popular search crawlers but misses crawler bots that specifically collect data for AI model training. In many contexts, applications of these models to predict people's private information may be based on shaky scientific grounds [2, 217], to put it mildly. However, such applications may still contribute to serious harms and discrimination regardless of their accuracy as long as some people are convinced of their predictive power and employ them to marginalize others. AI doxing may also take other forms such as reverse-image-search [14], a functionality supported by some search engines that may be abused for stalking and harassment. In turn, exploitative materials produced with GenAI (such as deepfake revenge porn or CSAM) might be amplified.

7.2.1.4   Consequence: Innovation Decay

Generative AI may find innovative new applications in information access. However, the excitement around these technologies and the significant investments from industry, government, and academia on corresponding research and development have broader implications for IR research. Next, we discuss some of the mechanisms associated with the research and development of generative AI that may potentially throttle innovation in information access technologies.

Mechanism: Industry Capture   The compute and data moat that concentrates power in the hands of big tech, as discussed earlier in Sect. 7.2.1.2, also creates significant barriers to entry for academic research. These barriers limit academic AI research to a handful of institutions that have the necessary means and connections to industry who provide access to compute and data resources to incentivize research in areas of their economic interests. Academics who want to contribute to research on large-scale AI systems or critique their sociotechnical impacts are pressured to play well with institutions holding monopolistic control over compute, data, and systems [150]. Access to "open-access" models—without the compute and data necessary to build them from scratch—allows academic researchers to invest in finding more effective applications of these technologies that serve industry interests, but not to reimagine/rearchitect them to in radically different ways. Students and other academics who may someday want to work in industry are shepherded into integrating themselves into this homogenized research agenda.

Such "industry capture" [237] allows for inordinate influence of the sociotechnical imaginaries[13] of profit-driven corporations over, for example, academic researchers [146]. This can thwart research that may not be immediately monetizable or challenges the status quo of power concentration and complements the "regulatory capture" by bigger tech companies [13, 130, 182]. As Mitra [146] asks:

---

[13] Jasanoff and Kim [103] define *sociotechnical imaginaries* as "collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology."

"*Whose sociotechnical imaginaries are granted normative status and what myriad of radically alternative futures are we overlooking?*" Narratives of the inevitability of these technologies that are hyped up to be both transformative forces for society and simultaneously posing existential risks for humanity (often purported by the same actors) only bolster their imagined importance to accumulate increasing global investments, including from governments. Researchers who care about sociotechnical impact and ecological sustainability are busy with enumerating the harms of rapidly emerging new AI technologies and chasing potential mitigations instead of having the full means to imagine and develop systems for social good. While industry practitioners can contribute to both identifying new research challenges grounded in real-world systems and practical methods to mitigate some of the risks of emerging technologies, it is imperative that we create avenues for increasing independent research while preserving the benefits of various modes of industry-academia collaborations.

Even as the grounded risks from these technologies (such as those discussed here) gather consensus from academic communities and civil society, it can be difficult to create space for alternative ways of development that are perceived as "slowing down." Critical research on sociotechnical harms of AI is also under risk when attempts are made to shift attention from concerns about real harms to marginalized people today to unsubstantiated imagined future concerns [71, 72]. Calls for regulations to address these imagined future harms [73] further detract from real progress and contribute to reinforcement of monopolistic powers of those who have already added these technologies to their arsenals. This has led some sociotechnical researchers in AI to explicitly draw attention to how these systems shift power (e.g., [23, 70, 107, 142]) and to prioritize research guided by alternative visions for sociotechnical futures grounded in universal emancipation and social justice [146]. It is thus important that access to investments to enable development is also available to those trying to not only mitigate existing systems' harms but also develop new avenues, including work on social good and new business models.

As generative AI starts to accumulate the lion's share of research investments, it may starve out other areas of information access research. Generative AI has had exciting but limited deployments in information access systems today. There are significant open challenges to making these models broadly useful, including but not limited to concerns of potential sociotechnical harms. There is a risk that if these challenges are not mitigated in spite of the extensive resources already invested on them at present, there may be calls for even larger investments in future prompted by the sunk cost fallacy.[14] It would be astute for the IR community to consciously continue to invest in research on systems and applications that societies need beyond what existing AI technologies make plausible [146, 184].

Mechanism: Pollution of Research Artefacts Risks to academic research from generative AI may also emerge through the applications of generative AI models

---

[14] https://en.wikipedia.org/wiki/Sunk_cost#Fallacy_effect

in IR scholarship—e.g., for authoring scientific papers and peer reviewing. There is evidence that researchers in computational sciences are already leveraging these tools [127], sometimes with hilariously terrible outcomes [163]. While the use of language models for light editing may (eventually) fall within the norms of socially acceptable behavior in research, their application in scholarship does raise concerns of plagiarism and scientific inaccuracies. This is an area that currently has more questions than answers, and the IR community would benefit from proactively considering potential implications of this trend on future IR research.

7.2.1.5   Consequence: Ecological Impact

Another important consequence of generative AI is its impact on the environment. In this context, it is important for us to consider the direct environmental cost of developing and deploying generative AI systems at scale as well as the potential impact of these technologies on the climate change discourse online.

Mechanism: Resource Demand and Waste   The ecological cost of deep learning models has been a subject of much concern and debate in the AI community [16, 17, 25, 56, 108, 161, 162, 199, 241]. Similar concerns have also been raised within the IR community with respect to the application of these models for information access [181, 249]. By some estimates, the computing power being utilized for deep learning research has been doubling every 3.4 months since 2012 [32]. In the USA, data centers consumed more than 4% of the total national electricity in 2022, and that number is projected to grow to 6% by 2026 [87]. Another study [15] estimates that by 2040, the information and communications technology industry on the whole will account for 14% of global emissions. Beyond emissions, data centers' water consumption is also raising alarm bells [50, 81, 84, 86, 89, 124, 152, 173]. By 2027, global AI demand may be responsible for withdrawal of 1.1–1.7 trillion gallons of fresh water annually [89, 124]. Serious concerns also revolve around the rising levels of electronic waste [112]. Even as we make progress in reducing the ecological cost of training and deploying the current AI models, we risk encouraging the development of even larger models and their wider deployment worsening the overall ecological impact (i.e., Jevons paradox).[15]

Mechanism: Persuasive Advertising   Generative AI may not only negatively impact the environment through increasing demand for natural resources and increasing generation of waste but may also supercharge climate change disinformation [43, 55, 68, 174, 175, 195]. For example, the fossil-fuel industry may attempt to sway public opinion through advertising that leverages generative AI's persuasion capabilities discussed in Sect. 7.2.1.2. Persuasive advertising may also be employed by other environment-unfriendly business models like fast fashion [42].

---

[15] https://en.wikipedia.org/wiki/Jevons_paradox

While the direct ecological cost of generative AI justifiably garners lots of attention, its potential impact on related online discourse also deserves scrutiny.

7.2.2   Risks

We categorize the risks of generative AI broadly to our society, to IR research, and to the environment. We map the first three consequences discussed earlier in this section—i.e.: (i) Information ecosystem disruption (Sect. 7.2.1.1) (ii) Concentration of power (Sect. 7.2.1.2) (iii) Marginalization (Sect. 7.2.1.3)—and their corresponding mechanisms as potentially contributing to the risks to society. We further map the last two consequences—i.e.: (iv) Innovation decay (Sect. 7.2.1.4) (v) Ecological impact (Sect. 7.2.1.5)—to the risks to IR research and the environment, respectively.

7.2.2.1   Risks to Society

Information access is a critical need of any democratic society and a necessary ingredient for social transformation [44, 78, 80, 96, 168]. It is also a social determinant of economic progress [151, 245] and health [147]. Disruptions to the information ecosystem bear potentially grave risks to most aspects of our social lives. A confluence of the pandemic [35, 183, 209], rising global conflicts [210, 213], and escalating climate catastrophes [102, 160, 169] is pushing the world toward precarious instability. Our information ecosystems are already struggling under the weight of misinformation and disinformation that in this critical moment is eroding public trust in online platforms, institutions, and each other. It is imperative that researchers and developers of information access systems prioritize safeguarding social interests and be vigilant in considering potential risks of disruption and ecosystem collapse when integrating generative AI technologies in the IR stack. This includes identifying the necessary conditions under which these technologies can be safely deployed and developing practical safeguards and alternatives.

Risks to society are not just from potential disruptions of the information ecosystem but also from how these technologies simultaneously concentrate power away from those at the margins of society. As institutions that develop and operate these technologies are themselves beneficiaries of this concentration, we need democratic oversights. If technologies further exacerbate already worsening wealth and power inequities, this additionally may pose severe threats to democratic institutions and human rights. There is an opportunity cost of not re-imagining information access in light of sociotechnical ambitions of human emancipation, culture, and knowledge production, instead of being constrained solely by what these emerging technologies make plausible and the homogenized visions put forth by institutions who wield these technologies [146].

7.2.2.2   Risks to IR Research

IR research can suffer from a confluence of different factors including the distancing of academic researchers from the data and compute they need to do their work and how narratives about the inevitability of AI technologies shapes what computational research gets funded. The concentration of access to the networks around these technologies in a subset of institutions shapes what is considered "foundational" or even "AI." Research on generative AI should not be performed only in the context of corporate economic interests while academia is hollowed out and prevented from exploring radical new methods that challenge the status quo. This risk of homogenization of academic research agendas and the opportunity cost of not exploring more diverse approaches to online information access can have material consequences. Instead, the IR community must be empowered with both the space and the resources necessary to explore a diversity of these visions and critique dominant narratives. IR research should have a plurality of work, which includes work with access to industry to change current practices. However, we especially also need to ensure that not all IR research is simply an extension of industrial system development and risk the demise of fundamental research on alternative avenues.

7.2.2.3   Risks to the Environment

Information access provides one of the large-scale application settings for generative AI. However, the impact of such wide-scale deployment of these technologies on the impending climate crisis should be a critical consideration. Climate costs pose substantial existential risks for ecosystems and people, in more direct ways than some other "existential risks" that lack adequate scientific basis but have nonetheless been popular discourse in some parts of the AI community. This means both choosing what to deploy and investment in methods to mitigate negative impacts that build on existing environmental work. As we discussed in Sect. 7.2.1.5, these concerns include not just the ecological cost of developing and deploying generative AI technologies but also their impact on online discourse on societal priorities.

7.3   Methods to Evaluate Risks and Impact

7.3.1   Evaluating the Impact of Generative IR Applications

Evaluating the impact of generative IR applications requires methods, as do data-informed interventions to steer that impact. Creating an LLM-based demo has become exceedingly easy. Understanding the impact of a system when it gets used in real-life contexts, and getting to a high-quality experience for a wide

variety of users, is much harder. Standards for impact assessment have not kept up a similar pace as tech developments. Khlaaf points out the need to carefully consider the differences in value alignment of the goals of a system and safety considerations, harms, and risks [113]. A wide range of online, offline, and human-assisted evaluations are possible—and necessary—to get a full sense of the impact of a system.

There are a number of frameworks that can provide helpful starting points for evaluating the impact of generative IR applications and potential quality or safety improvements. Not surprisingly however, they can measure quite different aspects of a system and its underlying models. Distinctions have to be made between evaluating a model, a system, or a technology as a whole. For example, standards for foundation model evaluations might not take into account the impact of a system that uses such a model (or a combination of models) in a specific application context.

Measurement and interventions are possible at every stage of the development life cycle of products and their underlying models and data. In this regard, general insights around, for example, harm mitigation interventions being possible throughout the machine learning life cycle [201] also apply to generative IR. To improve quality and safety, we need to be able to operationalize and measure the impact of potential interventions. This includes evaluations on aspects of that might be both system performance issues but also of societal importance, e.g., harmful/toxic output, hallucination, and differing model performance across languages/demographics.

7.3.2 Threat Identification, Assessment, and Modeling

When the emergence of a new technology or application becomes apparent, the assessment of whether this poses risks or opportunities within specific domains poses a challenge. Before development of a system, threats and opportunities can be identified. As Kapoor et al. [109] point out, it is crucial not to evaluate the risks and impact of new systems in isolation but rather in comparison with existing technologies. For example, the impact of usage of foundation models in *search* should be compared to existing Web *search*. For this purpose, Kapoor et al. present an evaluation framework that focuses on marginal risks, applied to Open Foundation Models. Their framework is based on threat identification work from cybersecurity and consists of six steps necessary to demonstrate such marginal risk. These steps are (1) threat identification, (2) evaluating existing risk absent open foundation models, (3) considering existing defenses absent open foundation models, (4) evidence of marginal risk of open foundation models, (5) ease of defending against new risks, and (6) outlining uncertainty and assumptions. Note that this framework does not set exact assessment criteria but rather defines the steps to get to such evaluations.

In practical settings, this might mean having to select standards for the development process (e.g., emerging standards from organizations such as the National

Institute of Standards and Technology (NIST) [3] or International Standards Organization (ISO) [12], company-specific standards such as Microsoft's Responsible AI Standard v2 General Requirements [143], or following new (local) legal requirements). However, mapping out potential consequences and identifying mechanisms that introduce risks in the specific context of a system needs to go much further. How to disrupt potential negative mechanisms in order to mitigate those risks requires gauging a wide range of consumer-side impacts [61] but also wider societal impacts. That includes frameworks focused on worker consequences [70] or practical methods focused on reducing the (legal) risks of using certain types of copyrighted or restricted training data vs. expected performance gains [144].

7.3.3 Evaluation During Model Development

7.3.3.1 Model Benchmarks vs. Actual System Context

LLM benchmarks are widely used to compare the quality and safety progress made by new *model* releases, resulting in model leaderboards on different scenarios. The Stanford HELM [198] leaderboard, for example, shows the performance of different LLM models on benchmarks, and these benchmarks include societal impact and bias-related measures. Their HELM ("holistic framework for evaluating foundation models") framework [126] uses scenarios and measures seven metrics. Those are accuracy, calibration, robustness, efficiency, and also more social impact-oriented fairness, bias, and toxicity. Each scenario focuses on one use case and consists of a dataset of instances, such as the LegalBench set of legal reasoning tasks [85] or medical board exam problem sets [104]. The larger BIG-bench ("Beyond the Imitation Game benchmark") [18] consists of 200+ tasks, contributed by hundreds of authors at a variety of institutes. More specific benchmarks for trustworthiness such as DecodingTrust, in turn, focus on subsets such as toxicity, stereotyping, adversarial and out-of-distribution robustness, privacy, machine ethics, and fairness [230], while, for example, the much more specific recurring TREC Fair Ranking track competitively evaluates systems according to how fairly they rank documents on a specific test task [60].

Paradoxically, while these benchmarks include aspects of societal impacts such as bias and toxicity, they do not necessarily cover the aspects that matter most in a specific application context in practice. Benchmarks are generally geared toward structured comparisons between models, *not* toward evaluating end-user applications in practice. This means that they may not be particularly suitable for a specific application and the people involved in its usage. In addition, using such large benchmarks can be quite resource-intensive, making "lite" versions necessary that are less comprehensive. Both HELM and BIG-bench are also implemented as Lite versions. However, the evaluation differences that arise from specific, lighter implementations of benchmarks can significantly impact model comparison results [196]. This makes it necessary to go beyond these benchmarks and ensure suitable

evaluations for the application at hand to avoid deriving conclusions about safety or responsibility devoid from actual application concerns.

7.3.3.2 Combining IR and Generative AI Evaluation Metrics

It is challenging that standards for measuring societal impact, including bias, fairness, etc., are yet scarce in IR *product* settings. For example, Smith et al. [193] provide an overview of different metrics available for evaluating bias and fairness in recommendation systems and the challenges practitioners face when choosing between them. In some cases, it may be more appropriate to, for example, focus on "traditional" performance and accuracy metrics but study the performance and subsequent quality of experiences for different groups of people by segmenting/slicing results by group. This approach assumes the ability to define relevant groups or relies on more advanced methods to find clusters that may—or may not—have significant differences in performance or quality.

   Specific methods might also be necessary to match new techniques. For example, Retrieval-Augmented Generation (RAG) might be used to include more reliable information in a specific domain and reduce hallucinations in an LLM setting. However, RAG does not necessarily fully solve every hallucination-related issue. Specific frameworks that fit an application context are still necessary to evaluate these techniques and their actual impact on aspects such as factuality within that context. One example is Saad-Falcon et al. [178], who present an evaluation framework, ARES, for RAG-assisted question-and-answering settings. This framework uses three evaluation scores: context relevance of the retrieved information, answer faithfulness (the answer's grounding in the retrieved context), and answer relevance to the question asked. These are similar to IR evaluations but might need adjustment to the setting at hand, and datasets used need to reflect actual needs in current circumstances.

7.3.3.3 LLMs to Evaluate LLM

Beyond specific metrics, ongoing research is investigating the efficacy of LLMs to evaluate LLMs (*LLM-as-judge*) [187, 232]. For example, [232] et al. use an LLM to rate the factuality of a long-form response to prompts while also using Google Search. While promising, such more complex evaluation constellations also lead to additional complexity in understanding what is being evaluated and changes therein as the evaluator LLM changes. This leads to having to validate the validation in itself [187]. While a human-and-LLM agent collaboration can help in this validation (as in, e.g., [187]'s EvalGen approach), the evaluation criteria cannot be fully separated from observation of model outputs, resulting in a feedback loop from output to adjusted evaluation criteria.

7.3.4 Evaluation Pre-/Post-system Release

7.3.4.1 Online Evaluation Using Actual User Behavior vs. Offline Evaluation

Whether evaluations are done online or offline can deeply impact results. Offline evaluations—even when using thoughtful standards—might not reflect what actual end users do in real-life settings or system performance over time. Online evaluations similarly are limited to which metrics have been instrumented and how actual user interactions are captured. It involves field testing, getting an IR system online and out to actual users, and analyzing their interactions with the system. It can include methods such as controlled experiments or extended A/B testing and analysis of interactions. Hoffmann provides an overview of the most common techniques used in IR settings [100].

7.3.4.2 Stress Testing, Red Teaming, and Qualitative End-User Evaluations

Beyond metrics and quantitative analysis–oriented methods, it is crucial to apply a combination of safety/security-inspired methods, user design, and User Experience (UX) research methods to understand the actual reactions of users.

The logistics around red teaming can provide a good glimpse into the importance of appropriate combinations of methods. Red teaming is a common way to test LLM applications for undesirable system responses [135, 145]. Red teaming can be automated using, for example, sets of (generated) prompts or done in full by human red teamers, including both the general public and invited experts. Using LLMs as red teamers [164] by generating risky prompts at scale, or using large-scale human red teaming efforts with thousands of participants who need access points, might yield different results. Human red team approaches in which "a group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture" (if we follow the definition from NIST) also lead to questions about tooling, recruiting, and operational process design. Markov et al. [135], for example, provide a helpful discussion of practical data challenges in content moderation use cases. In turn, model characteristics might have consequences on red teaming results. Ganguli et al. [69], for instance, find that RLHF models are increasingly difficult to red-team as they scale, while they do not find similar challenges for other models. Interestingly, this means that techniques such as RLHF that are explicitly meant to help align agents with human preferences could also result in challenges in evaluating the systems that use them.

This means that like any evaluation method, red teaming has to be combined with other types of stress testing, assessment of security issues, as well as evaluation of experiences of actual users. Khlaaf points out the need for carefully considering

what methods and terminology are appropriate for evaluations that probe for vulnerabilities of a specific system toward the outside world [113].

7.3.5 Societal Impact of a System Beyond Its Direct Implementation and Use

The impact of a system can reach much beyond its direct usage context. For example, the increasing demand for data and compute power of LLMs has environmental impact. However, such indirect impact can be hard to calculate without deep expertise. It is crucial to spend the time to evaluate evaluations methods for their suitability. Methods have been developed in both the IR and LLM communities around reducing environmental harm [181], and sustainability industry teams exist to ensure more energy-efficient data centers for both environmental and monetary reasons. Others in turn try to assess whether LLMs could help in generating more green code and develop metrics to assess the code's "green capacity" based on earlier sustainability metrics [216].

Similarly, a plethora of work points out the potential of amplifying and entrenching power structures through the usage of generative AI methods or changing market conditions through releasing new models for free [176], de facto changing standards to the model that gets used most in practice. However, IR and Machine Learning (ML) evaluation methods are not generally suitable for the analysis of such impact that a particular technique or system might have. Methods from political analysis and behavioral economics might be more suitable but are generally not shared in IR or ML venues. Challenging in the evaluation of systems is a deeper understanding of the long-term incentives that are created and the resulting "rational" use of LLMs in undesirable ways. A compounding challenge is that new incentives are also necessary to ensure that interventions from actual practice can be shared. Trust and safety teams might be doing scenario planning or prepare for incidents and crises.

7.3.6 Sharing Evaluation Methods

From the above selection of methods, which is by no means comprehensive, it is clear that practitioners have to carefully pick and choose which methods work for them. However, different organizations come from different evaluation traditions.

Incentives to share methods and results might not align with practical product team incentives and pressures. Metrics and standards for evaluations from actual practice are often not shared in scientific literature. Security community-style (external) red and (internal) blue teams, trust and safety incident monitoring approaches, IR communities' existing offline and online user feedback methods, or UX product testing approaches might be more (or less) top of mind, depending

Table 7.2 Different types of existing evaluation frameworks relevant for generative IR impact and safety. Note that this is not an exhaustive overview but rather a quick peek at the variety of methods evaluators can (and have to) choose from

Evaluation focus	Examples
Marginal system impact, e.g., release decisions in comparison with existing technology	Kapoor et al., risk framework based on cybersecurity [109]
Comparison benchmarks between LLM models that include fairness, bias, toxicity-type aspects	Benchmarks used in leaderboards, e.g., HELM [198], BIG-bench [18], or trustworthiness benchmarks [230]
Online or offline IR metrics, including accuracy or quality across groups	Online IR-evaluation methods [99], impact/fairness/bias metrics in recommendation systems [192]
Evaluation metrics using automated evaluation for specific LLM techniques or risks	LLMs as agents evaluating factuality of other LLMs' statements [187, 232]
Qualitative evaluation including human adversarial testing	Red teaming[69, 164] and UX evaluation

on the organization and prior expertise. This means there is a gap in the generative IR literature in terms of shared understanding of actual practices and efficacy of methods [48]. If we as a community are to properly address the social risks as outlined in Sect. 7.2.2.1, it is imperative we find fast and effective ways to share these methods and align them with practical needs, especially with the increasing speed of the field, the variety of fields involved, and volume of new techniques (Table 7.2).

7.4 Actors, Incentives, and Ways of Getting Organized

7.4.1 Incentives Toward Misuse of AI

Emerging AI capabilities and their consequences (good or bad) are a hot topic of discussion. But it is just as important to talk about incentives or why individuals or organizations might choose to use AI in certain ways.

Below are some examples of types of actors and their possible incentives that can lead to harmful uses of AI, along with ways in which some of them can be shifted in a more positive direction. AI can be transformative for human experience and quality of life, but only if incentives (both short term and long term) for its use are aligned with the benefits to humanity.

Actor State actors and ideological groups.

Incentive Geopolitical influence in favor or against something. This includes the use of extra-persuasive [238], micro-targeted content and deepfakes to sow malicious narratives [191], undermine support and trust in democratic institutions [148], weaken social cohesion, etc.

Modification The most effective way to modify this behavior is by making it prohibitively expensive or inconvenient to use AI for these purposes, through harsh legal consequences, content moderation, or counter-speech. The burden of implementing countermeasures falls on governments, content platforms, and community organizations.

Actor Criminal or unscrupulous organizations.

Incentive Financial gains from scams, ad-monetized Web site traffic, or product sales. This includes more legit-looking phishing content [229] and "Nigerian prince" letters or gaming search engines via AI-generated SEO-friendly content [156].

Modification The incentives for financial gain are always going to exist and be exploited; protection against them can take the form of better (AI-enhanced) cybersecurity and anti-spam tools, implemented and deployed by most consumer-facing Web surfaces.

Actor Commercial enterprises.

Incentive Economic competitive advantage and increased shareholder value. Taken to its worst extreme, this incentive can lead to deceptive or discriminatory business practices, hasty deployment of cheaply developed AI to customers [65], premature restructuring of teams [121], etc. In the case of social media platforms, the high engagement on polarizing or sensationalist content can lead the platforms to tolerate, encourage, and algorithmically amplify it.

Modification The same drive for competitive advantage can also be a force for good, particularly when it is aligned with public opinion or customer sentiment. The best-case scenario is when trustworthy and safe AI makes products more usable, attracting more customers (akin to Apple's "It just works" aesthetic that has no shortage of fans despite being more expensive than the competition). Government-led compliance requirements can also create positive incentives, like for food or car safety, and in some cases, a punitive legal strategy also works, like in the suing of tobacco companies or opiate producers, creating incentives for surviving companies to behave better.

Actor Individuals.

Incentives Faster completion of work tasks, improved social status, revenge against perceived slights, or exploitation of the vulnerable. At worst, these can lead to cheating, misrepresentation of one's identity of accomplishments, slander, deepfake pornography, or AI-enhanced grooming.

Modifications While some of these behaviors are illegal or fundamentally antisocial (and should be prosecuted as dictated by law), the urge to improve one's work performance or social status can be a good thing. If AI tools are designed to enhance human productivity while rewarding our creative impulses, and feel fun, joyful, and satisfying to use, people will be more likely to employ them to good ends.

7.4.2 Who Can Shift Incentives and How

In the broadest sense, it will take a whole-of-society approach to ensure that technological advances will align with the best interests of humans impacted by them (see Fig. 7.1). Technology builders (company and individual), governments, academia, and civil society all bear responsibility for ensuring that technological advances in information access align with societal interests. The rest of this section focuses on what can be done at the intersection of these groups or actors, since inter-group coordination is most often where things go awry.
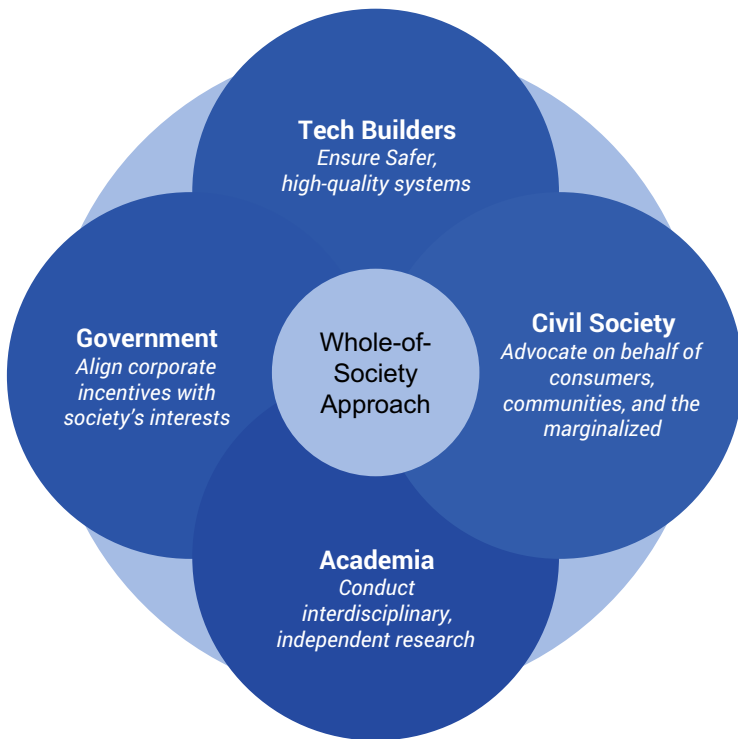

Fig. 7.1 Primary actors responsible for aligning technology with societal interests

7.4.2.1 Organizational Factors

While most of the literature and education in computer science by definition focuses on technical approaches, the impact of generative IR techniques can be influenced in other ways as well.

Changing work processes *within* organizations can have a direct impact on the expectations set on teams. This includes policies, explicit Go/No-Go procedures, roles and responsibilities to monitor systems, algorithmic impact assessments and model cards, or other types of documentation. In different organizations, the responsibility for different measurement and mitigation might look very different. In one organization, a machine learning team may be expected to look at the energy consumption of their system design choices, whereas other organizations might have a technical sustainability team. In another organization, a trust and safety or integrity team might deliver evaluations of system output toxicity, whereas in another organization, a separate data science team or product teams themselves might have to do this work. In any case, if this responsibility is unclear, it is much harder to get this work done.

External engagement can help address internal deficiencies. Especially for audiences working on generative IR systems, some of these might not necessarily be familiar routes. Examples include:

- **External advice and safety boards.** These are increasingly created by companies to provide external advice for more complex safety or content moderation questions. These include Facebook's Oversight Board,[16] which provides independent rulings on content moderation questions, parent company Meta's Safety Advisory Council,[17] or Spotify's Safety Advisory Board.[18] These do not necessarily have decision-making power but provide a more formalized way to advise external organizations and researchers.
- **Regulatory advisory groups and expert consultations.** Organizations such as the UN, EU, various regions, and countries working on future AI policy have all formed advisory boards (e.g., the UN AI advisory board,[19] the Nordic AI advisory board). Apart from such official avenues, individual lawmakers and legal firms often consult experts. While regulatory capture is a very real concern [236], this also allows for actually implementable regulation. This means however that considering the potential overlap between advisory boards, as well as perhaps a lack of overlap with more specific AI experts, not all relevant expertise will be represented.
- **Professional organizations.** Organizations such as Association for Computing Machinery (ACM), Institute for Electrical and Electronics Engineers (IEEE),

---

[16] https://www.oversightboard.com

[17] https://www.facebook.com/help/222332597793306/

[18] https://newsroom.spotify.com/2022-06-13/introducing-the-spotify-safety-advisory-council/

[19] https://www.un.org/en/ai-advisory-body

Association for the Advancement of Artificial Intelligence (AAAI), or the Trust & Safety Professional Association allow for formal and informal exchange of best practices. A major challenge is ensuring that best practices in fast-moving areas are also gathered and exchanged *between* organizations and to the public at large.

For the above arrangements, getting to collections of concrete examples of what has worked in the past is increasingly important. AI developments are speeding up, and increasingly diverse professional communities are both being impacted and getting involved. This makes efficient and effective coordination even more important. For policymakers, governmental agencies, and journalists, it may be hard to get an overview of which professional communities can provide actionable advice—especially with new AI developments being "louder" than, for example, long-standing IR communities. Inside of companies, in order to benefit from external advice or research, tech teams still have to navigate how to best work with external organizations. Researchers and non-governmental organizations in turn have to know where to invest their time and expertise most effectively and how to offer actionable advice to appropriate individuals or teams in tech companies. This includes big-picture scenario planning of where to best invest and how to create incentives that truly will have a positive impact. Implicit hierarchies of the value of different types of produced knowledge (e.g., "being the first" or "more technically complex"), but also a simple lack of knowledge about how certain processes work, can stand in the way of sharing of paved paths toward desired results and of sharing these in accessible ways. It can also involve very pragmatic on-the-ground work, such as knowing how to set up contractual arrangements that work for all parties (not a skill commonly taught in IR or AI-related programs).

7.4.2.2 Data-Focused Methods

While a complete overview of all different mechanisms to positively affect AI development is outside the scope of this chapter, one area does provide ample inspiration. Extensive literature exists on data labor and the need to understand how to effectively advocate for that labor's value [11, 75, 123, 218, 223]. Especially in the realm of training data concerns, multiple practical routes already exist, including:

- **Business and partnership model development,** including developing new types of licensing and new types of business partnerships [9, 190], along with ways to get funding to data creators. There is also research on the efficacy of suggested mechanisms, such as data dividends that are suggested as a means of AI profit sharing [227].
- **Collective action**. When new business models do not work out, coordinated action is imperative. These can be focused on data through data strikes [226], as well as large-scale labor organizing and strikes focused on treatment of data workers. More recently, the Hollywood strikes illustrated how those particularly impacted by the ways their work and likeness can be used as data can effectively

organize, lay out clear demands, and succeed through both technical and organizational competence. This included understanding what incentives are at play and what leverage data producers have [222]. Methods include data strikes to withhold data [205, 226] and data poisoning [54] techniques such as NightShade [94, 186], Glaze [185] and Mist [125]. Ways to empower end users and the wider public in their relationship with tech companies are important [228], as is understanding their potential leverage and means for protest through adjusted usage [122].

For effective research-informed mitigations, however, it is crucial that generative IR researchers have access to ways to learn how to effectively organize and navigate organizational and political structures or how to communicate their results to others. Implicit hierarchies in what knowledge is appreciated in generative IR circles can become a hurdle in effectively identifying and addressing the risks outlined in earlier sections, Sects. 7.2.2.1, 7.2.2.2, and 7.2.2.3. A critical factor is knowing which concrete situations matter, what to ask for in those situations, and how to assess whether impacts and risks are successfully steered.

7.5   Conclusion

In this chapter, we have presented a discussion on the sociotechnical implications of generative AI for information access. These deliberations are grounded in how these emerging technologies are currently being applied in IR applications as well as their future applications as being envisioned by practitioners and researchers. It is important to recognize that sociotechnical visions of what information access should look like in the future are not just shaped by what emerging technologies like generative AI make plausible but that visions for the future of information access in turn shape AI technologies themselves. Mitra [146] proposed the hierarchy of IR stakeholder needs shown in Fig. 7.2 and argued that IR research and system development require a fundamental shift toward re-centering societal needs and that we should reimagine information access as a vehicle for alternative futures. When contemplating the implications of emerging technologies, we risk of falling in the trap of limiting ourselves to how the technology (and its process of development) is today, rather than how it can be or *should be* in the future. Neither generative AI nor its application in the context of information access is predetermined. So while it is important that we consider potential harms of contemporary applications of generative AI in the context of information access, we close with some open question for the reader: *If not this status quo, then what—and especially how? What is the future of information access that we want to imagine for our collective well-being, and how can generative AI be another tool in the toolbox toward that transformation?*
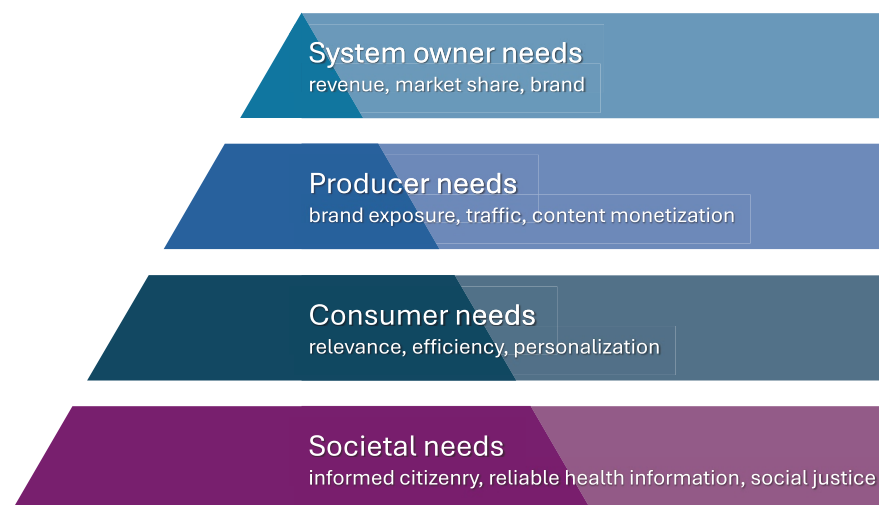
Fig. 7.2 Mitra's [146] hierarchy of IR stakeholder needs. More critical needs are at the bottom of the pyramid. This figure has been reproduced from the original paper with permission

References

1. Abid, A., Farooqi, M., Zou, J.: Persistent anti-Muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306 (2021)
2. Agüera y Arcas, B., Mitchell, M., Todorov, A.: Physiognomy's new clothes (2017). https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a
3. AI standards, NIST (2024). https://www.nist.gov/artificial-intelligence/ai-standards
4. Ai used to target kids with disinformation (2023). https://www.bbc.co.uk/newsround/66796495
5. Al-Sibai, N.: The top google image for Israel Kamakawiwo'ole is AI-generated. Futurism (2023)
6. Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoohi, A., Baraniuk, R.G.: Self-consuming generative models go mad (2023). arXiv preprint arXiv:2307.01850
7. Altenried, M.: The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. Capital Class **44**(2), 145–158 (2020)
8. Anguiano, D., Beckett, L.: How Hollywood Writers Triumphed Over AI–and Why It Matters. The Guardian (2023)
9. AP, OpenAI agree to share select news content and technology in new collaboration, 2023. https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/
10. Appel, G., Neelbauer, J., Schweidel, D.A.: Generative AI has an intellectual property problem. Harvard Bus. Rev. **7** (2023)
11. Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., Weyl, E.G.: Should we treat data as labor? moving beyond "free". In: AEA Papers and Proceedings, vol. 108, pp. 38–42. American Economic Association, Nashville (2018)
12. Artificial intelligence (AI) standards. https://www.iso.org/sectors/it-technologies/ai

13. Asokan, A.: UK government warned of AI regulatory capture by big tech. BankInfoSecurity (2024)
14. Baio, A.: 'most disturbing AI site on Internet' can find every picture of you that exists. Indy100 (2024)
15. Belkhir, L., Elmeligi, A.: Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J. Clea. Prod. **177**, 448–463 (2018)
16. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021)
17. Berreby, D.: As use of A.I. soars, so does the energy and water it requires. Yale Environment 360 (2024)
18. BIG-bench authors: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Trans. Mach. Learn. Res. (2023). ISSN 2835-8856. https://openreview.net/forum?id=uyTL5Bvosj
19. Birhane, A.: Algorithmic colonization of Africa. SCRIPTed **17**, 389 (2020)
20. Birhane, A., Cummins, F.: Algorithmic injustices: towards a relational ethics (2019). arXiv preprint arXiv:1912.07376
21. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184 (2022)
22. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: a case study of African-American English (2016). arXiv preprint arXiv:1608.08868
23. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp (2020). arXiv preprint arXiv:2005.14050
24. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
25. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models (2021). arXiv preprint arXiv:2108.07258
26. Brewster, J., Arvanitis, L., Sadeghi, M.: Funding the next generation of content farms: some of the world's largest blue chip brands unintentionally support the spread of unreliable AI-generated news websites (2023). NewsGuard
27. Brewster, J., Fishman, Z., Xu, E.: Funding the next generation of content farms: some of the world's largest blue chip brands unintentionally support the spread of unreliable AI-generated news websites (2023). NewsGuard
28. Burke, K.: 'biggest act of copyright theft in history': thousands of Australian books allegedly used to train AI model. The Guardian (2023)
29. Burke, K.: Generative AI is a marvel. is it also built on theft? The Economist (2024)
30. Burtell, M., Woodside, T.: Artificial influence: an analysis of AI-driven persuasion (2023). arXiv preprint arXiv:2303.08721
31. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
32. Cantrell, T.: The True Cost of AI Innovation. Scientific Computing World
33. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650 (2021)
34. Carroll, M., Chan, A., Ashton, H., Krueger, D.: Characterizing manipulation from AI systems. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–13 (2023)
35. Centers for Disease Control and Prevention et al.: CDC museum COVID-19 timeline. 2022, 2022
36. Chayka, K.: Is A.I. Art Stealing from Artists? The New Yorker (2023)

37. Chesterman, S.: Good models borrow, great models steal: intellectual property rights and generative AI. Policy Soc. puae006 (2024)
38. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
39. Coffey, D.: Māori are trying to save their language from big tech. Wired UK (2021)
40. Cohan, W.D.: Ai is learning from stolen intellectual property. It needs to stop. The Washington Post (2023)
41. Coldewey, D.: Thousands of authors sign letter urging AI makers to stop stealing books. TechCrunch (2023)
42. Coleman, J.: AI's climate impact goes beyond its emissions. Scientific American (2023)
43. Corbett, J.: Report warns generative AI could turbocharge climate disinformation. Common Dreams (2024)
44. Correia, A.M.R.: Information literacy for an active and effective citizenship. In: White Paper prepared for UNESCO, the US National Commission on Libraries and Information Science, and the National Forum on Information Literacy, for use at the Information Literacy Meeting of Experts, Prague, The Czech Republic (2002)
45. Couldry, N., Mejias, U.A.: Data colonialism: Rethinking big data's relation to the contemporary subject. Television New Media **20**(4), 336–349 (2019)
46. Cox, J.: Google news is boosting garbage AI-generated articles. 404 Media (2024)
47. Coyle, J.: In Hollywood writers' battle against AI, humans win (for now) (2023)
48. Cramer, H.: Practical routes in the UX of AI, or sharing more beaten paths. Interactions **29**(5), 89–91 (2022). ISSN 1072-5520. https://doi.org/10.1145/3555834
49. Crawford, K.: The trouble with bias. In: Conference on Neural Information Processing Systems, Invited Speaker (2017)
50. Criddle, C., Bryan, K.: Ai boom sparks concern over big tech's water consumption. The Conversation (2024)
51. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F.: The power of noise: redefining retrieval for rag systems (2024). arXiv preprint arXiv:2401.14887
52. del Rio-Chanona, M., Laurentsyeva, N., Wachs, J.: Are large language models a threat to digital public goods? Evidence from activity on stack overflow (2023). arXiv preprint arXiv:2307.07367
53. Dhawan, S.: Universities leveraging AI detectors: international students fear they may be wrongly accused of cheating. Financial Express (2023)
54. Dickson, B.: What is machine learning data poisoning? The Verge (2020)
55. Disinformation, Climate Action Against: Artificial intelligence threats to climate change (2024)
56. Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A.S., Smith, N.A., DeCario, N., Buchanan, W.: Measuring the carbon intensity of AI in cloud instances. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1877–1894 (2022)
57. Dupré', M.H.: Sports Illustrated published articles by fake, AI-generated writers. Futurism (2023)
58. Dupré', M.H.: Top Google result for "Edward Hopper" an AI-generated fake. Futurism (2023).
59. Duran, G.: The tech baron seeking to "ethnically cleanse" San Francisco. The New Republic (2024)
60. Ekstrand, M.D., McDonald, G., Raj, A., Johnson, I.: Overview of the trec 2021 fair ranking track. In: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings (2022)

61. Ekstrand, M.D., Beattie, L., Pera, M.S., Cramer, H.: Not just algorithms: strategically addressing consumer impacts in information retrieval. In: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part IV, pp. 314–335. Springer, Berlin, Heidelberg (2024). ISBN 978-3-031-56065-1. https://doi.org/10.1007/978-3-031-56066-8_25

62. El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., et al.: A mechanism-based approach to mitigating harms from persuasive generative AI (2024). arXiv preprint arXiv:2404.15058

63. Elish, M.C.: Moral crumple zones: cautionary tales in human-robot interaction (pre-print). Engaging Science, Technology, and Society (pre-print) (2019)

64. Ferrara, E.: Should ChatGPT be biased? Challenges and risks of bias in large language models (2023). arXiv preprint arXiv:2304.03738

65. Fowler, J.A.: Turbotax and h&r block now use AI for tax advice. It's awful. The Washington Post (2024). https://www.washingtonpost.com/technology/2024/03/04/ai-taxes-turbotax-hrblock-chatbot/

66. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. **30**(3), 411–437 (2020)

67. Gabriel, I., Ghazavi, V.: The challenge of value alignment: from fairer algorithms to AI safety (2021). arXiv preprint arXiv:2101.06060

68. Galaz, V., Metzler, H., Daume, S., Olsson, A., Lindström, B., Marklund, A.: AI could create a perfect storm of climate misinformation (2023). arXiv preprint arXiv:2306.12807

69. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Hume, T., Jacobson, J., Johnston, S., Kravec, S., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Joseph, N., McCandlish, S., Olah, C., Kaplan, J., Clark, J.: Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned (2022)

70. Gausen, A., Mitra, B., Lindley, S.: A framework for exploring the consequences of AI-mediated enterprise knowledge access and identifying risks to workers. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (2024)

71. Gebru, T.: Effective altruism is pushing a dangerous brand of 'AI safety' (2022)

72. Gebru, T., Torres, É.P.: Eugenics and the promise of utopia through artificial general intelligence. First Monday (2023)

73. Gebru, T., Bender, E.M., McMillan-Major, A., Mitchell, M.: Statement from the listed authors of stochastic parrots on the "AI pause" letter (2023). https://www.dair-institute.org/blog/letter-statement-March2023/

74. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N.A.: Realtoxicityprompts: evaluating neural toxic degeneration in language models (2020). arXiv preprint arXiv:2009.11462

75. Gershgorn, D.: GitHub's automatic coding tool rests on untested legal ground. The Verge (2021)

76. Gertner, J.: Wikipedia's moment of truth. The New York Times Magazine (2023)

77. Gillespie, T.: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press (2018)

78. Goldstein, S.: Informed Societies. Facet Publishing (2020)

79. Gonen, H., Goldberg, Y.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proceedings of the NAACL, pp. 609–614 (2019)

80. González, M.: A better-informed society is a freer society (2021). https://www.unesco.org/en/articles/better-informed-society-freer-society

81. Gordon, C.: AI is accelerating the loss of our scarcest natural resource: water. Forbes (2024)

82. Gordon, A.D., Negreanu, C., Cambronero, J., Chakravarthy, R., Drosos, I., Fang, H., Mitra, B., Richardson, H., Sarkar, A., Simmons, S., et al.: Co-audit: tools to help humans double-check AI-generated content (2023). arXiv preprint arXiv:2310.01297

83. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90 (2023)
84. Guerrini, F.: AI's unsustainable water use: how tech giants contribute to global water shortages. Forbes (2023)
85. Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D.N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G.M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J.H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., Li, Z.: LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models (2023)
86. Gupta, J., Bosch, H., Vliet, L.V.: AI's excessive water consumption threatens to drown out its environmental contributions. The Conversation (2024)
87. Halper, E.: Amid explosive demand, America is running out of power. The Washington Post (2024)
88. Hao, K.: Artificial intelligence is creating a new colonial world order. MIT Technology Review (2022)
89. Hao, K.: AI is taking water from the desert. The Atlantic (2024)
90. Hao, K., Hernández, A.P.: How the AI industry profits from catastrophe. MIT Technology Review (2022)
91. Hao, K., Seetharaman, D.: Cleaning up ChatGPT takes heavy toll on human workers. The Wall Street Journal **24** (2023)
92. Hardin, G.: The tragedy of the commons. In: Classic Papers in Natural Resource Economics Revisited, pp. 145–156. Routledge (2018)
93. Hays, K., Barr, A.: AI is killing the grand bargain at the heart of the web. 'we're in a different world.'. Business Insider (2024)
94. Heikkila, M.: This new data poisoning tool lets artists fight back against generative AI. MIT Technology Review (2023)
95. Hern, A.: TechScape: How cheap, outsourced labour in Africa is shaping AI English (2024). https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt
96. Higgins, S., Gregory, L.: Information Literacy and Social Justice: Radical Professional Praxis. Library Juice Press (2013)
97. Hoel, E.: Here lies the Internet, murdered by generative AI. The Intrinsic Perspective (2024). https://www.theintrinsicperspective.com/p/here-lies-the-internet-murdered-by
98. Hoel, E.: A.I.-generated garbage is polluting our culture. The New York Times (2024)
99. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Found. Trends Inf. Retr. **10**(1), 1–117 (2016). ISSN 1554-0669. https://doi.org/10.1561/1500000051
100. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Found. Trends®Inform. Retrieval **10**(1), 1–117 (2016). ISSN 1554-0669. https://doi.org/10.1561/1500000051
101. Hom, K.-L.: Rage baiting. Westside Seattle (2015)
102. IPCC, Climate Change et al.: The physical science basis, the working group i contribution to the UN IPCC's Fifth Assessment Report (wg1 ar5) (2013)
103. Jasanoff, S., Kim, S.-H.: Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication Of Power. University of Chicago Press (2015)
104. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl. Sci. **11**(14) (2021). ISSN 2076-3417. https://doi.org/10.3390/app11146421
105. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)

106. Kak, A., West, S.M.: AI Now 2023 landscape: Confronting tech power (2023). https://ainowinstitute.org/2023-landscape
107. Kalluri, P., et al.: Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature **583**(7815), 169–169 (2020)
108. Kanungo, A.: The green dilemma: can AI fulfil its potential without harming the environment? Earth.Org (2023)
109. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., et al.: On the societal impact of open foundation models (2024)
110. Kasirzadeh, A., Gabriel, I.: In conversation with artificial intelligence: aligning language models with human values. Philos. Technol. **36**(2), 1–24 (2023)
111. Khanal, S., Zhang, H., Taeihagh, A.: Why and how is the power of big tech increasing in the policy process? The case of generative AI. Policy Soc. puae012 (2024)
112. Khattak, R.: The environmental impact of e-waste. Earth.Org (2023)
113. Khlaaf, H.: Toward comprehensive risk assessments and assurance of AI-based systems. Trail of Bits (2023)
114. Klein, N.: AI machines aren't 'hallucinating'. but their makers are. The Guardian (2023)
115. Knibbs, K.: Scammy AI-generated book rewrites are flooding amazon. Wired (2024). https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/
116. Knibbs, K.: Your kid may already be watching AI-generated videos on YouTube. Wired (2024) https://www.wired.com/story/your-kid-may-be-watching-ai-generated-videos-on-youtube/
117. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. **110**(15), 5802–5805 (2013)
118. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of The ACM Collective Intelligence Conference, pp. 12–24 (2023)
119. Kugel, S., Hiltner, S.: A new frontier for travel scammers: A.I.-generated guidebooks. The New York Times (2023). https://www.nytimes.com/2023/08/05/travel/amazon-guidebooks-artificial-intelligence.html
120. LaFrance, A.: The rise of techno-authoritarianism. The Atlantic (2024)
121. Landymore, F.: Sports Illustrated lays off journalists after announcing pivot to AI content. Futurism (2023). https://futurism.com/the-byte/sports-illustrated-lays-off-journalists-ai-content
122. Li, H., Vincent, N., Tsai, J., Kaye, J., Hecht, B.: How do people change their technology use in protest? Understanding. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–22 (2019)
123. Li, H., Vincent, N., Chancellor, S., Hecht, B.: The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1151–1161 (2023)
124. Li, P., Yang, J., Islam, M.A., Ren, S.: Making AI less" thirsty": uncovering and addressing the secret water footprint of AI models (2023). arXiv preprint arXiv:2304.03271
125. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: preventing painting imitation from diffusion models via adversarial examples. arXiv preprint arXiv:2302.04578 (2023)
126. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C.D., Ré, C., Acosta-Navas, D., Hudson, D.A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S.M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., Koreeda, Y.: Holistic evaluation of language models (2023)

127. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., et al.: Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews (2024). arXiv preprint arXiv:2403.07183

128. Liesenfeld, A., Dingemanse, M.: Rethinking open source generative AI: open washing and the EU AI Act. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1774–1787 (2024)

129. Limbong, A.: Authors push back on the growing number of AI 'scam' books on Amazon. National Public Radio (2024). https://www.npr.org/2024/03/13/1237888126/growing-number-ai-scam-books-amazon

130. Liu, L.: Letter: setting rules for AI must avoid regulatory capture by big tech. Financial Times (2023)

131. Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., Liu, Y.: Prompt injection attack against LLM-integrated applications (2023). arXiv preprint arXiv:2306.05499

132. Liu, X., Yu, Z., Zhang, Y., Zhang, N., Xiao, C.: Automatic and universal prompt injection attacks against large language models (2024). arXiv preprint arXiv:2403.04957

133. Lonergan, R.: Mr. Justice Brandeis, great American. Mr. Justice Brandeis, Great American: Press Opinion and Public Appraisal (1941)

134. Möhlmann, M.: Algorithmic nudges don't have to be unethical. Harvard Business Review **22** (2021)

135. Markov, T., Zhang, C., Agarwal, S., Eloundou Nekoul, F., Lee, T., Adler, S., Jiang, A., Weng, L.: A holistic approach to undesired content detection in the real world. Proc. AAAI Conf. Artif. Intell. **37**(12), 15009–15018 (2023). https://doi.org/10.1609/aaai.v37i12.26752

136. Marr, B.: Is generative AI stealing from artists? Forbes (2023)

137. Martínez, G., Watson, L., Reviriego, P., Hernández, J.A., Juarez, M., Sarkar, R.: Towards understanding the interplay of generative artificial intelligence and the Internet (2023). arXiv preprint arXiv:2306.06130

138. Mathewson, T.: AI detection tools falsely accuse international students of cheating. The Markup (2023)

139. McMahon, C., Johnson, I., Hecht, B.: The substantial interdependence of Wikipedia and Google: a case study on the relationship between peer production communities and information technologies. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 142–151 (2017)

140. Mehdi, Y.: Bringing the full power of copilot to more people and businesses (2024). https://blogs.microsoft.com/blog/2024/01/15/bringing-the-full-power-of-copilot-to-more-people-and-businesses/

141. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: making domain experts out of dilettantes. In: ACM SIGIR Forum, vol. 55, pp. 1–27. ACM, New York (2021)

142. Miceli, M., Posada, J., Yang, T.: Studying up machine learning data: why talk about bias when we mean power? Proc. ACM on Hum.-Comput. Interact. **6**(GROUP), 1–14 (2022)

143. Microsoft responsible AI standard, v2, 2022. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl

144. Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N.A., Zettlemoyer, L.: Silo language models: Isolating legal risk in a nonparametric datastore (2023). arXiv preprint arXiv:2308.04430

145. Mishkin, P., Ahmad, L., Brundage, M., Krueger, G., Sastry, G.: Dalle-2-preview/system-card.md at main · openai/dalle-2-preview. https://github.com/openai/dalle-2-preview/blob/main/system-card.md

146. Mitra, B.: Search and society: Reimagining information access for radical futures (2024). arXiv preprint arXiv:2403.17901

147. Moretti, F.A., Oliveira, V.E.d., Silva, E.M.K.d.: Access to health information on the Internet: a public health issue? Revista da Associação Médica Brasileira **58**, 650–658 (2012)

148. Mularczyk, K.: Row over deepfake of Polish PM in opposition-party broadcast. Brussels Signal (2023). https://brusselssignal.eu/2023/08/row-over-deepfake-of-polish-pm-in-opposition-party-broadcast/

149. Muldoon, J., Wu, B.A.: Artificial intelligence in the colonial matrix of power. Philos. Technol. **36**(4), 80 (2023)
150. Murgia, M.: AI academics under pressure to do commercial research. Financial Times **13** (2019)
151. Mutula, S.M.: Digital divide and economic development: case study of sub-Saharan Africa. Electron. Library **26**(4), 468–489 (2008)
152. Naughton, J.: AI's craving for data is matched only by a runaway thirst for water and energy. The Guardian (2024)
153. Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory, and discussion. ACM J. Data Inform. Quality **15**(2), 1–21 (2023)
154. O'Gorman, M.: At the heart of artificial intelligence is racism and colonialism that we must excise. The Globe and Mail Web Edition (2023)
155. Oremus, W.: He wrote a book on a rare subject. then a ChatGPT replica appeared on Amazon. The Washington Post (2023) https://www.washingtonpost.com/technology/2023/05/05/ai-spam-websites-books-chatgpt/
156. Orland, K.: Lazy use of AI leads to Amazon products called "i cannot fulfill that request". Ars Technica (2024). https://arstechnica.com/ai/2024/01/lazy-use-of-ai-leads-to-amazon-products-called-i-cannot-fulfill-that-request/
157. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models (2023). arXiv preprint arXiv:2305.13661
158. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.: Automatic personality assessment through social media language. J. Pers. Soc. Psychol. **108**(6), 934 (2015)
159. Park, P.S., Goldstein, S., O'Gara, A., Chen, M., Hendrycks, D.: AI deception: a survey of examples, risks, and potential solutions (2023). arXiv preprint arXiv:2308.14752
160. Parmesan, C., Morecroft, M.D., Trisurat, Y.: Climate change 2022: impacts, adaptation and vulnerability (2022)
161. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training (2021)
162. Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D.R., Texier, M., Dean, J.: The carbon footprint of machine learning training will plateau, then shrink. Computer **55**(7), 18–28 (2022)
163. Pearson, J.: Scientific journal publishes AI-generated rat with gigantic penis in worrying incident. Vice (2024)
164. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G.: Red teaming language models with language models (2022)
165. Perrigo, B.: Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make Chat-GPT less toxic (2023). https://time.com/6247678/openai-chatgpt-kenya-workers/. Accessed 19 2023
166. Pierce, D.: You can now use the Dall-E 3 AI image generator inside Bing Chat. The Verge (2023)
167. Png, M.-T.: At the tensions of South and North: Critical roles of global South stakeholders in AI governance. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1434–1445 (2022)
168. Polizzi, G.: Information literacy in the digital age: why critical digital literacy matters for democracy. In: Informed Societies: Why Information Literacy Matters for Citizenship, Participation and Democracy, pp. 1–23 (2020)
169. Poynting, M., Rivault, E.: 2023 confirmed as world's hottest year on record (2024)
170. Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I.: A human rights-based approach to responsible AI (2022). arXiv preprint arXiv:2210.02667
171. Qi, X., Zeng, Y., Xie, T., Chen, P.Y., Jia, R., Mittal, P., Henderson, P.: Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv preprint arXiv:2310.03693 (2023)

172. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: predicting personality with Twitter. In: 2011 IEEE third international Conference on Privacy, Security, Risk and Trust and 2011 IEEE third international Conference on Social Computing, pp. 180–185. IEEE (2011)

173. Ren, S.: How much water does AI consume? The public deserves to know it (2023)

174. Report: Ai fueling climate change, energy usage and disinformation (2024). https://sustainablebrands.com/read/product-service-design-innovation/ai-fueling-climate-change-energy-disinformation

175. Report: Artificial intelligence a threat to climate change, energy usage and disinformation (2024). https://foe.org/news/ai-threat-report/

176. Robins-Early, N.: New GPT-4o AI model is faster and free for all users, OpenAI announces. The Guardian (2024)

177. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Mag. **36**(4), 105–114 (2015)

178. Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: Ares: an automated evaluation framework for retrieval-augmented generation systems (2024)

179. Sadeghi, M., Arvanitis, L.: Rise of the newsbots: Ai-generated news websites proliferating online. NewsGuard (2023)

180. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., Prabhakaran, V.: Re-imagining algorithmic fairness in India and beyond. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 315–328 (2021)

181. Scells, H., Zhuang, S., Zuccon, G.: Reduce, reuse, recycle: green information retrieval research. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2825–2837 (2022)

182. Schaake, M.: Big tech calls for 'regulation' but is fuzzy on the details. Financial Times (2021)

183. Scientist, N.: COVID-19: the story of a pandemic. New Scientist **10** (2021)

184. Shah, C., Bender, E.M.: Situating search. In: Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, pp. 221–232 (2022)

185. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23), pp. 2187–2204 (2023)

186. Shan, S., Ding, W., Passananti, J., Zheng, H., Zhao, B.Y.: Prompt-specific poisoning attacks on text-to-image generative models (2023). arXiv preprint arXiv:2310.13828

187. Shankar, S., Zamfirescu-Pereira, J.D., Hartmann, B., Parameswaran, A.G., Arawjo, I.: Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences (2024)

188. Shrivastava, R.: OpenAI and Microsoft sued by nonfiction writers for alleged 'rampant theft' of authors' works. Forbes (2023)

189. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R.: The curse of recursion: training on generated data makes models forget (2023). arXiv preprint arXiv:2305.17493

190. Shutterstock expands partnership with OpenAI, signs new six-year agreement to provide high-quality training data (2023). https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year

191. Simchon, A., Edwards, M., Lewandowsky, S.: The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus **3**(2), pgae035 (2024). ISSN 2752-6542. https://doi.org/10.1093/pnasnexus/pgae035

192. Smith, J.J., Beattie, L.: RecSys fairness metrics: many to use but which one to choose? arXiv preprint arXiv:2209.04011 (2022)

193. Smith, J.J., Beattie, L., Cramer, H.: Scoping fairness objectives and identifying fairness metrics for recommender systems: the practitioners' perspective. In: Proceedings of the ACM Web Conference 2023, WWW '23, pp. 3648–3659. Association for Computing Machinery, New York (2023). ISBN 9781450394161. https://doi.org/10.1145/3543507.3583204.

194. Solaiman, I.: The gradient of generative AI release: Methods and considerations. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 111–122 (2023)
195. Speare-Cole, R.: Generative AI could 'supercharge' climate disinformation, report warns. Independent (2024)
196. Srivastava, A., Rastogi, A., Rao, A., et al.: Beyond the imitation game: quantifying and extrapolating the capabilities of language models (2023)
197. Stahl, B.C., Eke, D.: The ethics of ChatGPT–exploring the ethical issues of an emerging technology. Int. J. Inform. Manag. **74**, 102700 (2024)
198. Stanford: Stanford HELM (2024). https://crfm.stanford.edu/helm/
199. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP (2019). arXiv preprint arXiv:1906.02243
200. Supporting open source and open science in the EU AI Act (2023). https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf
201. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '21. Association for Computing Machinery, New York (2021) ISBN 9781450385534. https://doi.org/10.1145/3465416.3483305.
202. Tacheva, J., Ramasubramanian, S.: Ai empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data Soc. **10**(2), 20539517231219241 (2023)
203. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models (2021). arXiv preprint arXiv:2102.02503
204. Tan, R., Cabato, R.: Behind the AI boom, an army of overseas workers in 'digital sweatshops'. The Washington Post (2023)
205. Tani, M.: New York Times drops out of AI coalition. Semafor (2023)
206. Taraborelli, D.: The sum of all human knowledge in the age of machines: a new research agenda for Wikimedia. In: ICWSM-15 Workshop on Wikipedia (2015)
207. Tarkowski, A.: How Wikipedia can shape the future of AI. Open Future (2023)
208. Tarkowski, A.: Stewarding the sum of all knowledge in the age of AI. Open Future (2023)
209. Taylor, L.: COVID-19: True global death toll from pandemic is almost 15 million, says WHO. Br. Med. J. **377**, o1144 (2022)
210. Taylor, A.: A historic rise in global conflict deaths suggests a violent new era (2023). https://www.washingtonpost.com/world/2023/06/29/conflict-war-deaths-global-peace-rise-casualty/
211. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences (2023). arXiv preprint arXiv:2309.10621
212. Thompson, B., Dhaliwal, M.P., Frisch, P., Domhan, T., Federico, M.: A shocking amount of the web is machine translated: insights from multi-way parallelism (2024). arXiv preprint arXiv:2401.05749
213. United Nations Meetings Coverage and Press Releases: With highest number of violent conflicts since Second World War, United Nations must rethink efforts to achieve, sustain peace, speakers tell security council (2023). https://press.un.org/en/2023/sc15184.doc.htm
214. Urman, A., Makhortykh, M.: The silence of the LLMs: cross-lingual analysis of political bias and false information prevalence in ChatGPT, Google Bard, and Bing Chat (2023)
215. Varghese, S.: How a Google search could end up endangering a life. iTWire (2021)
216. Vartziotis, T., Dellatolas, I., Dasoulas, G., Schmidt, M., Schneider, F., Hoffmann, T., Kotsopoulos, S., Keckeisen, M.: Learn to code sustainably: an empirical study on LLM-based green code generation (2024)
217. Vincent, J.: The invention of AI 'gaydar' could be the start of something much worse. The Verge **21** (2017)
218. Vincent, N.: Don't give OpenAI all the credit for GPT-3: You might have helped create the latest "astonishing" advance in AI too, 2020. https://www.psagroup.org/blogposts/62
219. Vincent, J.: The lawsuit that could rewrite the rules of AI copyright. The Verge **22** (2022)

220. Vincent, J.: Shutterstock will start selling AI-generated stock imagery with help from OpenAI. The Verge **25** (2022).
221. Vincent, J.: AI art tools stable diffusion and midjourney targeted with copyright lawsuit. The Verge (2023)
222. Vincent, N.: The WGA strike is a canary in the coal mine for AI labor concerns (2023). https://dataleverage.substack.com/p/the-wga-strike-is-a-canary-in-the
223. Vincent, N., Li, H.: GitHub Copilot and the exploitation of "data labor": A wake-up call for the tech industry (2021). https://www.psagroup.org/blogposts/62
224. Vincent, N., Li, H.: ChatGPT stole your work. so what are you going to do? (2023).
225. Vincent, N., Johnson, I., Hecht, B.: Examining Wikipedia with a broader lens: quantifying the value of Wikipedia's relationships with other large-scale online communities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)
226. Vincent, N., Hecht, B., Sen, S.: "data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In: The World Wide Web Conference, pp. 1931–1943 (2019)
227. Vincent, N., Li, Y., Zha, R., Hecht, B.: Mapping the potential and pitfalls of" data dividends" as a means of sharing the profits of artificial intelligence (2019). arXiv preprint arXiv:1912.00757
228. Vincent, N., Li, H., Tilly, N., Chancellor, S., Hecht, B.: Data leverage: a framework for empowering the public in its relationship with technology companies. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 215–227 (2021)
229. Violino, B.: AI tools such as ChatGPT are generating a mammoth increase in malicious phishing emails. CNBC (2023). https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html
230. Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S.T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., Li, B.: DecodingTrust: a comprehensive assessment of trustworthiness in GPT models (2024)
231. Warren, T.: Microsoft's new Copilot Pro brings AI-powered Office features to the rest of us, 2024. https://www.theverge.com/2024/1/15/24038711/microsoft-copilot-pro-office-ai-apps
232. Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., Le, Q.V.: Long-form factuality in large language models (2024)
233. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models (2021). arXiv preprint arXiv:2112.04359
234. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229 (2022)
235. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.S.: Challenges in detoxifying language models (2021). arXiv preprint arXiv:2109.07445
236. Whittaker, M.: The steep cost of capture. Interactions **28**(6), 50–55 (2021). ISSN 1072-5520. https://doi.org/10.1145/3488666
237. Whittaker, M.: The steep cost of capture. Interactions **28**(6), 50–55 (2021)
238. Williams, R.: Humans may be more likely to believe disinformation generated by AI. MIT Technology Review (2023). https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/
239. Williams, A., Miceli, M., Gebru, T.: The exploited labor behind artificial intelligence. Noema Mag. **13** (2022). https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/
240. World economic forum global risks report 2024 (2024) https://www.weforum.org/publications/global-risks-report-2024/

241. Wu, C.J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al.: Sustainable AI: environmental implications, challenges and opportunities. Proc. Mach. Learn. Syst. **4**, 795–813 (2022)
242. Xiang, C.: OpenAI used Kenyan workers making \$2 an hour to filter traumatic content from ChatGPT. VICE (2023)
243. Yom-Tov, E., Dumais, S., Guo, Q.: Promoting civil discourse through search engine diversity. Soc. Sci. Comput. Rev. **32**(2), 145–154 (2014)
244. Youyou, W., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. Proc. Natl. Acad. Sci. 112(4), 1036–1040 (2015)
245. Yu, P.K.: Bridging the digital divide: Equality in the information age. Cardozo Arts Ent. LJ **20**, 1 (2002)
246. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences (2019). arXiv preprint arXiv:1909.08593
247. Zuboff, S.: The age of surveillance capitalism. In: Social Theory Re-Wired, pp. 203–213. Routledge (2023)
248. Zuboff, S., Möllers, N., Wood, D.M., Lyon, D.: Surveillance capitalism: an interview with Shoshana Zuboff. Surveill. Soc. **17**(1/2), 257–266 (2019)
249. Zuccon, G., Scells, H., Zhuang, S.: Beyond CO_2 emissions: the overlooked impact of water consumption of information retrieval models. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 283–289 (2023)