Mapping the individual, social, and biospheric impacts of Foundation Models

Andrés Domínguez Hernández*
adominguez@turing.ac.uk
The Alan Turing Institute
London, UK

Shyam Krishna*
skrishna@turing.ac.uk
The Alan Turing Institute
London, UK

Antonella Maia Perini*
aperini@turing.ac.uk
The Alan Turing Institute
London, UK

Michael Katell
mkatell@turing.ac.uk
The Alan Turing Institute
London, UK

SJ Bennett
sj.bennett@durham.ac.uk
University of Durham
Durham, UK

Ann Borda
aborda@turing.ac.uk
The Alan Turing Institute
London, UK

Youmna Hashem
yhashem@turing.ac.uk
The Alan Turing Institute
London, UK

Semeli Hadjiloizou
shadjiloizou@turing.ac.uk
The Alan Turing Institute
London, UK

Sabeehah Mahomed
smahomed@turing.ac.uk
The Alan Turing Institute
London, UK

Smera Jayadeva
sjayadeva@turing.ac.uk
The Alan Turing Institute
London, UK

Mhairi Aitken
maitken@turing.ac.uk
The Alan Turing Institute
London, UK

David Leslie†
d.leslie@qmul.ac.uk
Queen Mary University of London
London, UK

ABSTRACT

Responding to the rapid roll-out and large-scale commercialization of foundation models, large language models, and generative AI, an emerging body of work is shedding light on the myriad impacts these technologies are having across society. Such research is expansive, ranging from the production of discriminatory, fake and toxic outputs, and privacy and copyright violations, to the unjust extraction of labor and natural resources. The same has not been the case in some of the most prominent AI governance initiatives in the global north like the UK's AI Safety Summit and the G7's Hiroshima process, which have influenced much of the international dialogue around AI governance. Despite the wealth of cautionary tales and evidence of algorithmic harm, there has been an ongoing over-emphasis within the AI governance discourse on technical matters of safety and global catastrophic or existential risks. This narrowed focus has tended to draw attention away from very pressing social and ethical challenges posed by the current brute-force industrialization of AI applications. To address such a visibility gap between real-world consequences and speculative risks, this paper offers a critical framework to account for the social, political, and environmental dimensions of foundation models and generative AI. Drawing on a review of the literature on the harms and risks of foundations models, and insights from critical data studies, science and technology studies, and environmental justice scholarship, we identify 14 categories of risks and harms and map them according to their individual, social, and biospheric impacts. We argue that this novel typology offers an integrative perspective to address the most urgent negative impacts of foundation models and their downstream applications. We conclude with recommendations on how this typology could be used to inform technical and normative interventions to advance responsible AI.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**; • **Applied computing** → **Sociology**; • **General and reference** → **General literature**.

KEYWORDS

risks, harms, foundation models, AI governance, responsible AI

ACM Reference Format:
Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, SJ Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. 2024. Mapping the individual, social, and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3630106.3658939

1 INTRODUCTION

Since the release of ChatGPT at the end of 2022, there has been considerable and sustained interest across policy, academia, and public discourses in the risks posed by artificial intelligence (AI), and particularly generative AI. Though ChatGPT may have generated the most column inches, its underpinning model—GPT 3.5 and,

*Equal contribution as lead authors
†Also with The Alan Turing Institute.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658939

subsequently, GPT 4—was just one example of wider developments in the area of so-called "foundation" or "frontier" models.

Foundation models are AI technologies trained on very large, "broad" datasets that can be applied to a wide range of tasks and purposes [25]. Their general, unspecified purpose and widespread deployment leads to concerns about potential unanticipated impacts when used in novel areas. These models are considered to form the "foundation" or base architecture of other systems. For example, a number of new applications such as ChatGPT and the conversational features of the Bing search engine, have been built on top of successive versions of OpenAI's GPT foundation model, which has been designed for natural language processing tasks.

While the techniques involved in developing and deploying foundation models are not new, the scale of the data used for training, the coordination of global networks of labor to process these massive amounts of data, the increased architectural complexity, and the development of increasingly powerful processors and other computational resources have together made possible new levels of predictive and generative sophistication. Due to the resources required to produce foundation models, the most prominent ones known to exist are the products of large and well-financed technology companies, such as Anthropic, Cohere, Hugging Face, Meta, OpenAI, Microsoft, and Alphabet (the latter two being substantial funders of other firms).

Responding to the rapid rollout and large-scale commercialization of foundation models, large language models (LLMs), and generative AI, an emerging body of work is exploring the myriad impacts these technologies are having across society. Such research is expansive. Prominent topics range from the production of discriminatory, fake, and toxic outputs and privacy and copyright violations to the unjust extraction of labor and natural resources. Foundation models present complex issues for law, policy, and practice, which arise concretely from the intertwined array of socio-technical systems in which their design, development, and use are embedded. These technologies pose risks and hazards that emerge from real-world contexts. Such issues arise from the global supply chains and labor sources that support their production, as well as the market forces and regulatory environments that influence their funding and financing. Risks likewise arise from the sociohistorical realities and legacies of inequity and exclusion that shape the data on which they are trained and from the patterns of privilege and socioeconomic stratification that influence the composition of the project teams that build them. All these as real-world contexts lead to the real-world consequences that have been the subject of much critical and responsible AI research on the impact of foundation models and generative AI over the past year. Yet, despite the importance of centering this empirically oriented work, much of the agenda-setting policy and public discourse emerging primarily from the main geographical centres of innovation on foundation models (and relatedly LLMs and generative AI) has focused on sensational, catastrophic, and largely speculative risks relating to extreme and hypothetical scenarios.

The narrative of catastrophic risks was amplified by much mainstream, English language, media coverage of foundation models throughout 2023, fueled by high profile statements declaring potential existential threats from AI [141, 142]. Most notably a statement published in May 2023, and signed by many prominent figures in the field of AI, claimed that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" [38]. While the research behind these claims is scant and broadly disputed, this narrative has reached far into predominant policy and regulatory discussions, including high profile international initiatives in the global north like the UK's AI Safety Summit [83] and the G7's Hiroshima AI Process [155]. Such processes, while presenting valuable opportunities to bring together the international community in discussing the extant risks that AI presents to people, society, and the planet, and corresponding approaches to needed regulation, have been dominated by technical discourses with narrow framings of "safety" and have prioritized discussions of speculative global catastrophic or existential risks of AI.

Concerns about this narrowing of AI policy and regulatory discourse have been echoed by various academic communities who have questioned the lack of "ideological and demographic diversity" of the AI safety field [116]; and the privileging of speculative narratives as effective distractions from existing and well-documented risks and harms of AI systems today [78, 92, 226]. A wealth of evidence demonstrates in particular the disproportionate negative impacts that AI systems have on marginalized communities—communities who are also underrepresented in decision-making processes regarding the design, development, deployment, or governance of AI. These issues, which are linked to historical and entrenched power asymmetries, are now being exacerbated by AI technologies, and risk becoming increasingly more urgent.

Therefore, through a review of the literature on the harms and risks of foundation models and generative AI, in this paper, we aim to address this visibility gap between real-world consequences and speculative risks. We propose a critical framework capable of adequately illuminating socio-technical risk contexts and attending to the social, political, and environmental dimensions of foundation models and generative AI. Through reviewing the existing evidence base our research maps a diverse set of risks and harms impacting individuals, communities, society, and the biosphere.

The paper starts with a discussion on the theoretical framework employed to investigate risks and harms in the context of foundation models. This framework abductively informs the methodological approach, which is detailed in section 3. Section 4 present our findings by delineating three distinct levels of algorithmic impacts—individual, social and biospheric. We conclude discussing how the integrative perspective we offer can aid in understanding the negative impacts of foundation models, and can help shape responsible AI futures.

2 THEORETICAL LENS: EXPANDING VIEWS ON ALGORITHMIC RISKS AND HARMS

There is a growing and multidisciplinary body of literature which explores the potential risks and harms posed by the relatively recent widespread popularization of foundation models. Although these risks and harms share many similarities with those presented by other types of AI, and indeed other data-intensive technologies more generally, it is crucial to examine how the proliferation and widespread adoption of foundation models may introduce new challenges, as well as further complicate existing ones.

The concept of risk is pervasive in modern societies and has become a central aspect of the AI governance discourse. Sociological studies have argued that theories of risk management and securitization emerged as a means to grapple with the uncertainties of modernity [18] and "making the future secure and certain" [186]. The understanding of risk and harm, as well as how they interrelate, varies greatly across domains. For legal, policy, organizational, or actuarial purposes, risk is typically approached through quantifiable measures and anticipatory models of potential adverse or harmful events such pandemics, financial losses, health and safety incidents, or operational disruptions [63, 147, 151, 170]. Risk is therefore not inherently an issue of morality and can be understood as a context-dependent and relational concept—that is, expressed in relation to how factors like technical developments, organizational structures, or innovation may adversely impact a desired future and be variously interpreted by different actors [234]. For example, consider a company which uses a job applicant screening software and has discovered that it makes racially biased screening recommendations. Where policy makers and civil society may be concerned about the various ways in which the technology undermines individual rights and social equity, a company might instead be more likely to focus on the risk of reputational damage and how this could in turn result in public distrust, low technology adoption, and commercial impacts.

Critical social science perspectives underline the multiple ramifications of identifying, analyzing, and managing risk associated with the impacts of technology in society [200, 244]. In this, some areas of research critically examine the various ways in which the risks and harms resulting from data and algorithmic systems are patterned across society at different levels [35, 133]. For instance, the field of science and technology studies (STS) has a well-established body of work exploring the interdependencies between technology and society, challenging the widespread deterministic view of technology as operationally autonomous and inherently value-neutral, and of innovation as following a predictable linear trajectory [136]. When it comes to examining the impacts of AI, scholars in the fields of STS, critical data studies, and socio-legal studies have proposed to conceptualize the harmful effects of data-intensive technologies as issues relevant to human rights, social justice, cultural institutions, the rule of law, the public sphere, economic systems, and the environment [51, 64, 137, 138, 196, 204, 243]. It is also from a critical perspective that some scholars increasingly highlight the need to approach algorithmic governance not only in terms of speculative notions of risk, but also in response to actual and observed harms to humans and the environment [73].

To build an understanding of risks and harms of technology from a moral and ethical standpoint, it is key to consider the questions of how power is distributed across the entire ecosystem of actors, where and how power manifests, and how asymmetries of power and information relate to differential impacts in society. Here it is helpful to draw insights from feminist and data justice scholarship, socio-legal studies, and the environmental justice literature. In particular, the concept of intersectionality makes explicit how social hierarchies manifest in different outcomes and experiences depending not only on the identity of the impacted individual or group, but also on the multiple compounding intersections of social identity and standing. These intersections go beyond crude constructions and classifications of individuals in society, to recognize the ways in which nuances of context, identity, and circumstance distinctly shape people's understanding and experiences of harm [52].

To capture these differential impacts as they relate to the proliferation of foundation models, in this paper we propose, following Leslie and Rossi [119], to assess risks and harms according to their individual, social, and biospheric levels of impact, contributing to an emerging dialogue within the FAccT and AI ethics communities [194, 239]. This perspective allows for a nuanced understanding of the observed and anticipated impacts of foundation models by recognizing the inherent socio-technical and globally entangled nature of these technologies. We use this framework to map the risks and harms found in the literature to date and offer an integrative perspective to address the most urgent negative impacts of foundation models and their downstream applications.

3 METHODS: SNOWBALL AND STRUCTURED SEARCH

We conducted a review of literature using two complementary and reinforcing qualitative literature review approaches which have been shown to be useful in studies aimed at surfacing patterns and trends in corpora of academic research [246].

We first applied a snowball approach [245], crowd-sourcing among all the co-authors, a core set of recent papers focused broadly on algorithmic harms and risks [19, 22, 25, 58, 105, 110, 192, 239]. These papers were selected based on our inclusion criteria of covering a broad range of multidisciplinary literature addressing risks and harms associated with foundation models. Populating a Zotero database, we followed snowball sampling approaches starting from the set of core articles and then carrying out a manual search using the references in those articles. Forward snowballing was used to identify new papers that focus on algorithmic risks and harms citing those in the core list using Google Scholar and Internet-based searches to access the abstract and full-text. We identified a total of 64 articles including journal articles, conference papers, and preprints drawing broadly on technical and socio-technical literature in which key risks and issues of foundation models, LLMs, and generative AI are described historically and contemporaneously. The search inclusion criteria covered English language publications and the most recent publication was of October 2023.

Complementing the snowball approach, we conducted a structured database search to identify primarily academic articles with a focus on the risks and harms associated with foundation models, LLMs, and generative AI.

Inclusion criteria covered English language journal articles, including conference papers and pre-prints, but excluding grey literature, monographs, commentaries, correspondences, and opinion pieces. The following electronic databases were searched: arXiv, ACM Digital Library, IEEE, Scopus, and Web of Science.[1] After the removal of duplicates and non-relevant papers, there were 114 papers in total for analysis, of which 11 were also papers in the snowball search results. The resulting set of papers were published

[1]The ACL Anthology database was not used as a source for the search given this database's core focus on computational linguistics and natural language processing.

in the period January 2018 to July 2023. The start date aligns with the date of appearance in the literature of transformer architectures which underpin LLMs, such as the BERT large language representation model [229].

We applied an abductive approach [213, 233] to analyze the combined corpus of literature 167 papers. The abductive approach was used to code the results and subsequently map key relationships between the risks and harms found in the literature guided by our theoretical framework described in Section 2. Firstly, we extracted applicable keyword information from the abstract of each primary study as part of an initial coding process. The resulting keywords served largely as summative and process attributes of a risk or type of harm relevant to foundation models or LLMs. Three researchers subsequently clustered these attributes and assigned them into parent categories. The codes were subsequently refined by all co-authors (see Appendix for search strategy). In this stage, we considered whether at least one of the risks and harms listed in the paper implied or directly referred to individual, social, or biospheric impacts to map the papers onto our three-level framework. These codified findings are further analyzed in Section 4.

3.1 Limitations

We note that our structured search showed a prevalence of preprint papers (non-peer review) in the result set, some of which had been cited extensively in both preprint and peer-review articles.[2] We used citations only as a broad indication of the preprints' impact. While some preprints were highly cited, these may have not undergone a rigorous peer-review process, and their findings should be interpreted with caution. In our sample, we analyze articles solely as an indicator of the range of concerns raised in the existing literature. The selected papers were also limited to the English language, and date range of January 2018 to July 2023. This range by default provided only a snapshot of a growing and multidisciplinary body of work which is arising in response to the fast development and roll-out of foundations models and LLMs, especially generative AI related studies. In addition, our structured literature search, which is not intended as systematic, only centred academic articles as a relevant account of the discourse on risks and harms of foundation models and LLMs to inform governance. We also note that this body of research is not focused exclusively on risks and harms, but often addresses, or points to, mitigations and governance of foundation models and AI at large. An analysis of other non-academic and informal sources was not part of this study and it warrants additional research. Notwithstanding, the codified mappings provide a substantive view of the research landscape and the range of concerns that are being raised around foundation models and LLMs, their development, implementation, and applications.

[2]We made a distinction between citations of preprints referenced by other preprints, some of which exceeded 50 citations, and those referenced by peer-reviewed publications. Those preprints that were more extensively cited had citations in both preprints and peer-reviewed articles.

4 MAPPING INDIVIDUAL, SOCIAL, AND BIOSPHERIC IMPACTS OF FOUNDATION MODELS

The result set of 14 risks and harms coding (Table 1) is telling of the wide-ranging and multifaceted challenges posed by foundation models and AI systems at large. Papers typically engage with three or more aspects, exemplifying the broad spectrum of concerns in the literature to date. For instance, Zhuo et al. [258] delve into potential biases, toxicity, and issues related to the reliability and robustness of ChatGPT. Notably, a substantial set of search results focuses on bias and discrimination. These are most typically related to LLMs, and include works such as van der Wal et al. [227] and Huang et al. [99], which address multiple biases and harmful stereotypes, and others, like Abid et al. [1], Felkner et al. [72], Ovalle et al. [159], and Gadiraju et al. [76] that concentrate on specific societal biases. The result set also reveals that papers that focus on specific or a limited number of risks and harms often scrutinize issues of unreliable performance of foundation models, or misinformation and propaganda. While other issues such as lock-in and opacity, or overdependence in human-computer interaction, are less prevalent in the result set, it is important to note that their absence does not imply that they are not identified as relevant within the existing literature. Instead, they are simply not the primary focus of most papers in our dataset.

In the following, we elucidate three distinct levels of algorithmic impact (individual, social, and biospheric) with examples stemming from the literature and categorize the risks and harms emerging from our search using this typology (see Table 1). This lens on algorithmic harm aims to account for the intersectional dimensions and socio-technical embeddedness of foundation models with the aim to better inform potential technical and normative interventions to advance responsible AI.

4.1 Individual risks and harms

It is now undeniable that AI systems can have adverse impacts to individuals which may be physical, psychological, or financial, but may also negatively affect a person's dignity, reputation, or fundamental rights and freedoms. Indeed, individual-level risks and harms arising from the implementation of data-intensive technologies have been largely documented in the literature, with individual impacts having immediate as well as long-lasting implications for those at the receiving end of such harms [176]. Equally so, a focus on individual rights has been at the core of conceptions of data protection, privacy, and security within digital regulation such as the European Union's General Data Protection Regulation (GDPR), where protection and redress mechanisms privilege notions of personal data or data subjects [196].

One of the biggest ethical concerns associated with AI systems has been their role in creating discriminatory, biased, and unfair outcomes which impact an individual's safety, health, wellbeing, and integrity. The proliferation of applications based on foundation models has not only augmented these concerns but has given rise to a host of new ethical, privacy, and safety issues. Biases and social prejudices present in training datasets and encoded in foundation models trickle down to applications like chatbots or image generators, thereby reproducing existing patterns of harm, exclusion, and

Table 1: A mapping of foundation models risks and harms according to their level of impact (individual, social, biospheric). The table shows the number of papers across categories of harm and levels of impact, and the percentage (in brackets) of the total number of papers in the set in which the risk/harm appears. Papers appearing in more than one category were counted more than once. See Appendix for an explanation of each risk and harm category.

Risks and Harms Category	Individual	Social	Biospheric
Bias and societal prejudices	71 (42.5%)	71 (42.5%)	0 (0%)
Misinformation, disinformation and propaganda	45 (26.9%)	45 (26.9%)	1 (0.6%)
Unreliable performance	43 (25.7%)	40 (24%)	0 (0%)
Cybersecurity risks and harms	37 (22.2%)	27 (16.2%)	0 (0%)
Privacy risks and harms	28 (16.8%)	21 (12.6%)	0 (0%)
Systemic social and economic risks and harms	18 (10.8%)	24 (14.4%)	1 (0.6%)
Legal and regulatory violations	17 (10.2%)	17 (10.2%)	0 (0%)
Environmental effects and ecological disruption	8 (4.8%)	8 (4.8%)	19 (11.4%)
Misuse	13 (7.8%)	14 (8.4%)	19 (11.4%)
Lock-in and opacity risks	13 (7.8%)	15 (9%)	0 (0%)
Overdependency in human-computer interaction	11 (6.6%)	6 (3.6%)	0 (0%)
Data risks and harms	7 (4.2%)	7 (4.2%)	0 (0%)
Value misalignment	2 (1.2%)	2 (1.2%)	0 (0%)
Extreme or catastrophic risks and harms	2 (1.2%)	2 (1.2%)	1 (0.6%)

discrimination [1, 76, 203]. For instance, chatbots—such as OpenAI's ChatGPT or Google's Bard—are capable of producing natural language responses to prompts through complex statistical associations [25, 254]. These responses are enabled by the vast amounts of data that the foundation models have been trained on, comprising mostly of internet-scraped datasets that have been shown to encode biased and harmful associations across protected categories of religion, disability, gender, race, and ethnicity [1, 61, 148], and also at the intersections of these characteristics [85, 211].

When it comes to issues of privacy and online safety, LLMs have been found to be effective tools for fraudulent activities such as scams, impersonation, and phishing, given their capability to generate personalized and highly convincing messages at scale [59, 89, 150]. Research has also shown that LLMs are prone to leaking private or sensitive information contained within their training data corpus [36, 106]. Because these models are trained on data scraped from the Internet, privacy violations might take place unbeknownst to impacted individuals thereby limiting their ability to consent or opt out [158]. Furthermore, models can associate different pieces of data to an individual, which poses privacy and security risks wherein sensitive information or even wrongful associations are revealed without consent [98]. Indeed, in some instances, chatbots have been used with the intention of extracting protected information and bypassing restrictions through the use of malicious methods like prompt hacking, jailbreaking [174, 236], and prompt injection [84, 164].

Lastly, the proliferation of foundation models complicates the already significant negative impacts digital technologies have on people's psychological wellbeing. In the last decade, researchers have raised concerns over how hyperconnectivity and unhealthy relationships with technology can have serious mental health implications and influence how people form individual identities and construct self-hood [5, 31, 134]. These issues are now exacerbated with the emergence of personalized AI chatbots capable of producing human-like responses. New research into the use of these conversational agents for mental health support warns about the potential for people's overdependence on these agents and the risk of social isolation [135]. Similarly, documented cases of users self-harming as a result of exposure to toxic interactions with AI agents evidence the devastating impacts these technologies can have on vulnerable individuals [117, 235, 249]. For example, prior research has demonstrated that individuals who are isolated may be more vulnerable to misleading information on social media platforms [26]. When individuals are subject to algorithmic nudging and personalized information campaigns, they are at risk of loss of autonomy, dignity, and integrity [180].

4.2 Social risks and harms

While the rise of foundation models has generated huge expectations about their positive transformational potential for society in many domains, there are also concerns about how their adoption can contribute to worsening existing social inequalities, biases, and injustices. Beyond individual risks and harms, there are different ways in which technology can have wider collective and societal impacts. One way to conceptualize collective risks and harms is to look at the compounded or aggregate impacts to a particular group, or to the functioning of a society as a whole [137].

Risks and harms to communities can be characterized by how particular demographics and group formations within society are distinctively impacted by technology, particularly when it comes to vulnerable groups with protected characteristics, e.g., migrants, refugees, LGBTQI+ communities, Muslims, the elderly. [3] As we have shown, not only can these technologies produce responses that are discriminatory and psychologically triggering for individual users, but the reproduction of hegemonic views, harmful stereotypes, and biases can have detrimental consequences at the collective level, affecting marginalized groups and communities including those who are not direct users, or are far removed from the technology. For instance, Abid et al. [1] demonstrate the persistent religious bias present in foundation models, noting that prompts containing the

[3]We note that notions of community are complex, often comprising loosely held webs of relational experience and knowledge, which share geo-spatial, identity-based, and/or digital spaces amongst other facets, with community membership subjective and fluid [32]. Here, communities are defined as groups sharing characteristics pertaining to intersectionality, power, and positionality [183].

word "Muslim" produced responses with violent language. These negative associations, in turn, can spur targeted hate and disinformation campaigns centered around anti-Muslim or Islamophobic sentiments.

Collective forms of harm can also be understood from a relational lens. The relational lens captures the adverse impacts to social relationships of interdependence, trust, and solidarity, thereby negatively impacting people's experiences of belonging, or their capacity to flourish through their relationship to others. A rich and interdisciplinary body of research has shed light on the range of interpersonal effects arising from people's interactions with algorithms, including the affective impacts of the algorithmic sorting of content on social media feeds [33, 67, 202] and the proliferation of misinformation as detrimental to the quality of the public sphere and democratic deliberation [20, 220, 225]. Foundation models in particular can increase the scale and speed at which disinformation campaigns can be disseminated across the information ecosystem [161, 198]. As generative AI applications powered by foundation models flood the public sphere with fake information, there is a risk of eroding public trust in the information that circulates online, further fueling social polarization and the creation of echo chambers [113, 195].

Lastly, algorithmic systems are inextricably embedded in wider social, political, and institutional structures, and as such they also have a bearing on those structures [243]. To grapple with the impacts that foundation models can have at a structural level, one must consider geopolitical, socioeconomic, legal, and organizational manifestations of power [119]. For instance, some scholars have warned about the tight concentration of power arising from the limited pool of actors controlling technological capabilities globally [50, 57, 157]; the influence of these power asymmetries on the development of international policies, standards, and regulations for technological practices [15, 43, 47]; and the overwhelmingly invasive data practices carried out by governments and companies [68, 74, 219].

The collective and deeper structural effects of these technologies are difficult to anticipate and measure as they manifest over time as AI becomes more widespread and embedded in society. Nonetheless, it is helpful to analyze and situate foundation models in the context of well-known social implications linked with digital technologies. For instance, going back to the issues related to the public sphere, one could argue that the effortless production and spread of targeted misinformation and propaganda enabled by foundation models, can have lasting effects on wider cultural, social, and political structures, namely established democratic processes, representation, participation in public life, and institutional trust. Given the evidence of public and private actors deploying technologies to generate manipulative content to sway public opinion [48, 75], scholars have warned about an increasing risk to social cohesion given the mounting erosion of trust in democratic structures [115].

4.3 Biospheric risks and harms

Lastly, biospheric risks and harms refer to the potential adverse impacts to ecological systems and environments, as well as wider impacts to the biosphere at the planetary level [169]. Importantly, this category also includes harms to the individuals and communities negatively impacted by environmental effects and ecological disruptions resulting from AI development and use. In other words, biospheric, social, and individual risks and harms are fundamentally intertwined. A good entry point to understand biospheric impacts is to look at the largely hidden socio-material entanglements of data-intensive systems. That is, the complex supply chains, labor, and underpinning materials crucial to AI infrastructures—including the Rare Earth Elements (REEs) that are necessary to manufacture the semiconductors essential to computer hardware. The global nature of the supply chains of foundation models pose individual, social, and environmental risks by entangling processes of extraction, knowledge production, and ultimately decision-making [177].

The development of supercomputers which utilize REEs, and enable foundation models to be trained, has been found to have significant material footprints globally [12]. The mining of silicon and other rare metals not only cause environmental harms, but can rupture local communities [4, 51] and cause individual harm to the workers mining these elements [241]. Furthermore, the negative consequences are usually borne disproportionately on vulnerable or marginalized communities largely located in the global south [34]. Thus, when put in the context of global infrastructures, the interrelated nature of these risks and harms becomes clearer. Robbins and van Wynsberghe argue that the "interconnectedness of AI with ecological, social, and economic systems" [179] pose very real risks regarding societal lock-in to AI infrastructures. That is to say that beyond a certain point, the nature of the supply chains and materials powering certain AI infrastructures cannot be changed, whilst ecological harms would continue to multiply because of resultant carbon emissions, REE mining, and other processes. Given these potential path dependencies, we are currently at a crucial juncture for developing ethical governance mechanisms that address the interlocking risks of the biospheric and human facets of AI ecosystems.

The development and implementation of data-intensive systems like foundation models is also known to have high local environmental costs in the form of energy consumption required to train increasingly large and complex machine learning systems [113, 130], and heavy water consumption needed for data center cooling [121]. The amount of compute needed to train large-scale models has doubled every 3.4 months since 2012 [9, 88], which translates to higher energy expenditures, and the use of costly resources even if the improvements in model accuracy are modest (i.e., diminishing returns in terms of model accuracy come at increased computational cost, see [187]). For example, the carbon emissions of training Google's BERT were roughly those of a transatlantic flight [207]. All of this at a time when curbing our global emissions is crucial to slowing climate change down and effectively mitigating its effects [222]. Efforts to develop responsible technologies that minimize their cascading impacts on the environment are hindered by the difficulty in comprehensively assessing carbon and energy footprints of training large models. Tools have emerged enabling practitioners and developers of these energy intensive technologies to monitor and track their models' emissions [11].

Although the environmental costs of AI are borne by everyone on the planet in terms of negative externalities contributing to climate

change and ecological deterioration, they are not uniformly distributed among the world's populations or regions. The allocations of benefits and risks replicate existing patterns of environmental injustice, coloniality, and "slow violence" [152], in which a disproportionate exposure to risks and harms is borne by marginalized communities; be these in terms of pollution, destruction of local ecosystems, or involuntary displacement. By contrast, the majority of the economic benefits reaped by the use of the model mostly go to the model's proprietary owner(s), even when these technologies adopt an open source ethos [126].

Finally, the impact of pollution and ecosystem degradation on interdependent and connected biological systems (including sentient non-human animals) cannot be ignored [45, 207]. AI has the potential to put non-humans at further risk (e.g., animals and the ecological systems in which they live) through surveillance, monitoring, and testing for which the benefits are largely directed to humans and society. Potential harms arise from data collection on animals in intensive factory farming to increase productivity, or data tracking of protected wildlife, which hunters and poachers can hack to illegally hunt or trade animals [29, 45]. Beyond these physical harms to animals, the functional use of AI in image recognition and as recommender systems demonstrate biases in differentiating species borne out from underlying social factors [86]. There is an urgent need to understand and respond in a systemic way to the interconnected and shared vulnerabilities of humans, animals, and the ecosystems in which we live and share [45].

4.4 Summary of findings

Our review shows that a large proportion of papers (40%) raise concerns about the potential for foundation models to perpetuate and amplify hegemonic views, harmful stereotypes, and societal and behavioral biases on an unprecedented scale. A similarly large set of papers (20%) also point to issues related to the creation and spread of misinformation and propaganda as well as the potential exploitation of these models for fraudulent services and cybersecurity attacks by malicious actors. Furthermore, the individual and societal impacts stemming from inconsistent or undesirable performance constitute a significant part of the mapped academic literature, making the case for cautious approaches to the deployment and use of such models.

Although less prevalent, a subset of the literature mapped point to concerns in connection to a range of technical and supply chain challenges of foundation models, and the potential compounding effects of those challenges as foundation models become more widespread. These range from the reliance on proprietary software and lack of transparency that enable lock-in and opacity risks to issues like outcome homogenization, narrowing of the market or monopolization, and the perpetuation of inequalities, which affect societal and economic structures. Our mapping also reveals that approximately 10% of the mapped papers address the environmental risks and harms associated with foundation models. While the primary focus of these latter works is on the biospheric-level impacts, some authors explore the interconnections between the adverse impacts of foundation models on ecology and animals and their consequential impacts on individuals and societies. This underscores the need for expanded exploration and attention to holistically understand the complexities surrounding the environmental adverse impacts of foundation models. Lastly, despite the extensive emphasis in much of the governance discourse, less attention is given in the academic literature to the extreme scenarios of existential, catastrophic, and other speculative risks of foundation models. These findings highlight the current visibility gap between real-world consequences and speculative risks, and sheds light on the areas requiring urgent and greater attention and efforts.

5 DISCUSSION: GRAPPLING WITH THE SCALE AND INTERCONNECTEDNESS OF FOUNDATION MODELS

Foundation models are highly consequential technologies which have sparked discussions about a spectrum of transformational impacts from the most promising to the most concerning. They pose unprecedented governance challenges but also offer an opportunity to draw lessons from, as well as re-examine, the landscape of socio-technical impacts of data-intensive technologies. While emergent and novel, foundation models do not arrive de novo; they are "built on an installed base" [201] and inherit the tools and methods of prior generations of machine learning and neural network technologies. As such, many of the foundation-model-related risks and harms that we highlight are likely to appear familiar because they occur, to some degree, in adjacent and precursor technologies.

Even so, two key differentiating characteristics of foundation models are their massive scale and widespread embeddedness. Foundation models comprise hundreds of billions of parameters, trained on mountains of data, that consume enormous resources for both training and deployment. In particular, the scale of foundation models means that the risks and harms they present are not only likely to be magnified and amplified, but that this will happen in ways which transcend national and political boundaries, requiring a multi-pronged and transnational response. Harms that may have been minimal or just minimally attended to in prior generations of technology are now made visible and urgent. For example, data-intensive systems have been criticized for some time for their carbon footprint [205], but the high visibility and widely publicized demands of foundation models have brought this issue to the fore to the point that it is unsurprising to see this aspect discussed in the technical literature [25].

Another differentiating characteristic of foundation models is their embeddedness. Foundation models are conceptualized and architected as the base models for many and diverse types of downstream applications. The embeddedness of foundation models renders them invisible yet pervasive. As a result of their platformized architecture, foundation models form the basis of many thousands of extensions, and as such, the negative impacts and harms stemming from foundation models may be obfuscated and rendered relatively intractable. These two characteristics—scale and embeddedness—position foundation models to be both highly adaptive, highly elusive, and highly dangerous. We argue that any assessment of risks and harms should account for these socio-technical interdependencies, and any design of mitigations and policy responses should be commensurate with the level of impact be it individual, societal, or biospheric. The value of our proposed framework for conceptualizing risks and harms is that by decomposing these into individual,

social, and biospheric impacts, we provide a conceptual tool with which to challenge attempts to narrow down salient risks and harms in a way that is meaningful only to some discrete set of affected actors and thereby limited in effect. The prevailing international policy discourse focusing on the technical safety of "frontier AI" systems, which has led to a flight from confrontation of the full spectrum of hazards presented here, is a signal example of such an erroneous narrowing [92]. In this paper, we underline the necessity of the opposite approach. When identified at their empirical sources, the risks and harms of foundational models become visible from all angles. They cut across individual, collective, and environmental levels, spreading over and affecting entire populations, including entire socio-material and biophysical ecologies of humans and non-humans. In this way, the impact and importance of the full range of risks and harms cannot be invisibilized or waved away in pursuit of economic, geopolitical, or other short-term goals.

5.1 Visibility gaps in the current assessment of algorithmic impact

As we have shown in this paper, the literature on risks and harms of foundation models is expansive and is likely to continue to grow. Grappling with such a vast and heterogenous landscape is a challenging task both for those who are attempting to build a nuanced understanding of the technology as well as for those who are seeking to inform the debate on mitigations and governance. We recognize notable attempts in the literature to taxonomize these risks [239], as well as significant progress being made internationally to reckon with the scale and embeddedness of foundation models by framing AI governance as a human rights and transnational issue [101, 221]. However, some of the most prominent governance initiatives, particularly in Europe and the US, has thus far fallen short of contending with the most problematic harms stemming from unequal patterns of data, labor, and resource extraction and instead has focused largely on risks to adoption, technical safety issues, and catastrophic risks.

One of the biggest challenges in assessing and anticipating algorithmic harm has to do with limited evidence and difficulties related to observing the indirect manifestations of harm, as well as foreseeing its effects over time. Grasping these complexities involves nuanced and context-dependent understanding. For instance, there are direct measurable impacts from the energy used to train or operate an LLM which are immediate and visible, but indirect harms arising from AI applications may only come to light over time or when enough evidence and research is made available. An AI system enabling the unsustainable extraction of mineral sites or unfair labor practices in the data labelling supply chain, are just a couple of illustrations of how negative indirect harms are likely to be largely hidden from view [154]. Direct harms to individuals may translate to indirect harms to society and vice versa. For instance, the potential harm that can ensue with data misuse or privacy breach (e.g., personal data) is not just limited to the individual who is directly manipulated, but indirectly affects the interests of society at large [196]. Where anthropocentric and species biases exist in AI, there are further examples of direct harm which remain understudied, largely because wider biospherical needs are not part of the conversation [45]. Algorithmic systems that ignore animals or privileges a particular view of animal welfare while ignoring others can exacerbate these consequential impacts, and not least can lead to indirect and direct harms to interconnected ecosystems shared by both humans and animals [29, 45].

5.2 Building a socio-technical scaffold to technical interventions

The current strategies for the mitigation of foundation model generated risks and harms that have most traction among policymakers and governments predominantly focus on technical interventions [156]. There is thus an urgent necessity to delve into the social context within which these interventions are situated. This is crucial, especially since the advent of generative AI represents not only a significant milestone in technical advancement but also simultaneously transforms the very fabric of far-reaching socio-digital infrastructure like the Internet and its social experience [90, 91].

Consider, for instance, the issue commonly framed in technical terms as "model collapse" or "data pollution". In this scenario, the generative output from the widespread experimental use of chatbots and image generators feeds into public data pools. This influx poses the risk of influencing future datasets that AI models will subsequently incorporate. The amalgamation of human and AI-generated content potentially undermines the quality and diversity of AI-generated outputs [139]. While this cycle undoubtedly necessitates technical intervention, the risk typology outlined in this paper enables the identification of concurrent impacts within the social sphere—impacts like the undermining of the long-term integrity of the information ecosystems on which modern democratic ways of life rest.

The integrative perspective we offer emphasizes, for example, that the risks surrounding the AI-enabled generation of biased or harmful images go beyond just offending individuals; in the aggregate, these systems have the potential to change social narratives around communities and to lock in cultural prejudices at scale, replicating and augmenting patterns of structural discrimination and injustice. Likewise, given the scale of their generative abilities, these technologies have broader planetary implications that derive from the cumulative costs of mass industrialized compute. Concentration on the perceived technical complexity of foundation models often mistakenly occludes such a clear ecological view of the social impact of generative outputs. Regardless, evolving research rightly insists that the AI community must grasp the subtleties, social contexts, and boundaries of human interaction with AI as a user-oriented technology, as well as the social and longitudinal aspects of innovation, more broadly [118, 173, 184]. The risk mapping presented in this paper facilitates this understanding. Even a singular, seemingly innocuous creation of politically incorrect content by an AI tool, a feature that many such tools still permit [87], can accumulate and result in a proliferation of societally effective bias within the community. This bias is amplified by the repetitive contributions from multiple individuals. Consequently, the framework presented here helps us understand how concerns usually detected or faced at individual levels, in fact, scaffold larger risks and contribute to higher-level concerns, thereby providing a deeper socio-technical understanding of foundation models and AI at large.

5.3 Toward an integrative perspective on risks and harms

Our framework illustrates concretely how claims that there may be gains from the implementation and use of foundation models that outweigh their potential and observed harms need to be examined and nuanced through sharpened socio-technical lenses. Such claims often rely on a strictly utilitarian calculation in which the overall potential "benefits" of foundation model application may outweigh the total harms and risks and are hence largely incapable of accounting for more systemic considerations. However, as shown by our framework, foundation model harms are frequently difficult to track and measure—and this affects the capacity to coherently weigh benefits and harms. For instance, harmful use cases that are relatively intelligible when measured individually could have impacts and consequences that are much harder to trace at social and biospheric levels. A moderate, but cumulative and difficult-to-perceive, harm to planetary health has implications for billions of people whereas a significant, but straightforward, harm to some individuals may be quite limited in scope. This makes utility calculations of benefits and harms difficult to perform with sufficient precision to capture their full range, scale, and scope. For example, Bommasani et al. [25] argue that the benefits of releasing large models—such as applications that translate text in otherwise underserved languages—outweigh the risks of misuse and abuse by malicious actors. Within their analysis, there is also a recognition that relatively few firms and organizations have sufficient resources and capacities to produce foundation models and that efforts to develop them for use by less-resourced non-elites is nascent and unlikely to produce models with similar capacities. The analysis, however, fails to perceive the deeper risks engendered by the asymmetrical power structures and dynamics that lead to these inequitable differentials in resources, capacities, and access.

Where a handful of highly self-interested and profit-driven companies control the data, compute, and skills infrastructures on which the development and use of foundation models rely, social harms arising from expanding inequality, wealth polarization, concentration of economic power, and privilege biases that lead to the escalating marginalization of minoritized groups will likely abound. Risk analyses that fail to acknowledge macro-scale issues like this will discount socially consequential adverse impacts. In this case, the very fact that the architecture of foundation models relies on high resource concentration should be a launching pad for expanding the narrow lens on technical risk to account for the ways in which foundation models play a role in reproducing harmful social hierarchies and planetary degradation. Such recognition is also crucial for contending with the potential for foundation models to lead to hard-to-reverse and long-term effects as this technology progressively becomes more embedded in society. What we call an integrative perspective is essential for overcoming the dominant utilitarian and performance-oriented approach in the AI governance discourse which has tended to frame societal challenges in terms of quantifiable trade-offs between risks and benefits.

6 CONCLUSION

In this paper we have confronted a blind spot in the evolving AI governance landscape that derives from its reliance, particularly among prominent policymakers in the global north, on speculative risks and selective seeing. Drawing on the rapidly growing multidisciplinary body of research on foundation-model-generated risks and harms, we have shown how discerning this through integrative and socio-technically curved lenses better discloses the full spectrum of impacts across individual, social, and biospheric levels. We have argued, in this respect, that there exists a visibility gap between the range of concerns and evidence raised in the critical, empirically anchored literature, and the abstract and mainly hypothetical issues focused on within some of the most influential international AI governance initiatives. Such a gap reflects the convergence of the power dynamics, private interests, and geopolitical priorities that have agenda-setting consequences in the AI governance ecosystem—which is a challenge to coherently grappling with the unprecedented industrialization of large scale data-driven technologies, rapidly transforming veritably every domain of life and communities around the world. While potentially beneficial and transformative, foundation models also pose numerous risks to people, society, and the planet. We have aimed to deepen understandings of this broad range of risks by bridging the technical aspects of foundation models with their socio-technical underpinnings, connecting individual concerns to collective and planetary issues, and doing justice to the multifaceted and differential impacts these models have on affected communities. Our conceptualization of risk, particularly regarding the potential transformations effected by these technologies, demonstrates that the visibility of risks and harms should not be concealed or obscured by speculative concerns about existential threats, hypothetically conceived. Instead, as under the framework presented here, an understanding of risks should be grounded in robust evidence that is observable at various levels, demonstrating the potential of these adverse impacts to escalate, over time, and to widen in their scale and scope. Ultimately, the framework we have presented enables a comprehensive assessment of algorithmic impacts for which an interdisciplinary dialogue is key. As such it can be applied as an analytical tool to inform socio-technical mitigations and to fundamentally expand existing toolkits for algorithmic fairness, transparency, and responsible AI.

IMPACT STATEMENT

1) *Description of the ethical concerns the authors mitigated while conducting the work (as part of an ethical considerations statement):* Throughout the process of developing the research questions and methodology, we foregrounded our work in an acute awareness and recognition of the role power plays in shaping conversation around the design, development, and deployment of technology. We grounded the evaluation of risks and harms as they relate to foundation models within well-established scholarship and critical discussion surrounding how power is distributed, the pervasiveness of stark power asymmetries, and how the differential impacts of foundation models –and technology at large—are experienced across multiple actors and layers of the ecosystem. We allowed this grounding to inform the framing of our research questions and our methodology, ensuring that a multidisciplinary body of scholarship was drawn on to inform our investigation. This approach included drawing on critical data studies, science and technology studies, and environmental justice scholarship, amongst

other fields. It is through this grounding that we aimed to challenge dominant discourses that adopt speculative perspectives on the risks and harms presented by the rapid deployment of foundation models; discourses which overwhelmingly distract from current, real-world consequences. We therefore aim to foreground our work in observed impacts. Through the proposed individual – social – biospheric framework, we work to acknowledge the inherently entangled and interdependent nature of socio-technical systems and their impacts, rather than further dominant discourses of fragmentation and division.

2) *Reflections on how their background and experiences inform or shape the work (as part of a researcher positionality statement):*
The final form and content of this work was shaped by a few factors. Firstly, all researchers involved in the development and writing of this work are academically trained in research, and are based at research and higher-education institutions in the Global North, where the primary language is English. As such, the research itself was limited to the English language. The researchers involved in this work, however, represent a variety of communities and come from a diverse range of backgrounds, including lived and research experience in the Global South. Throughout the process of developing the work, we continuously reflected on– and engaged with– how our own values, beliefs, perspectives, and lived experiences inevitably shaped the work presented here. The research team which developed this work is interdisciplinary, coming from a variety of research specializations including anthropology, sociology, data science, and information science, as well as diverse and global industry experience.

3) *Reflection on the adverse, unintended impact the work might have once published (as part of an adverse impact statement):*
In presenting this work, we are acutely aware that the research is squarely grounded in a rapidly evolving field. As such, we acknowledge that the proposed arguments and recommended pathways this research presents might not be relevant in the near to distant future, and we do not prescribe any solutions that could have an impact in the future. However, we also acknowledge that –despite these intentions—our risks and harms codes, and the subsequent categorizations we develop, may have unintended impacts; either through mis-categorizing certain risks and harms, or not accounting for others. These codes and categories were developed by our team's collective conceptualization of the terms 'risk' and 'harm'; as we are each working within our individual and group positionalities, there may be other forms of risks and harms not fully captured in this research.

ACKNOWLEDGMENTS

The authors wish to thank Claudia Fischer and Janis Wong for their key contributions to an earlier draft of this paper, as well as the three anonymous reviewers and area chair for their very constructive feedback. This work was supported by the Ecosystem Leadership Award under the EPSRC Grant EP/X03870X/1, the Arts and Humanities Research Council Grant AH/Z505584/1, and The Alan Turing Institute.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298 – 306. https://doi.org/10.1145/3461702.3462624 Type: Conference paper.

[2] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. http://arxiv.org/abs/2303.12528 arXiv:2303.12528 [cs].

[3] Ali Al-Kaswan and Maliheh Izadi. 2023. The (ab)use of Open Source Code to Train Large Language Models. http://arxiv.org/abs/2302.13681 arXiv:2302.13681 [cs].

[4] Rami Al Rawashdeh, Gary Campbell, and Awwad Titi. 2016. The socio-economic impacts of mining on local communities: The case of Jordan. *The Extractive Industries and Society* 3, 2 (April 2016), 494–507. https://doi.org/10.1016/j.exis.2016.02.001

[5] Anders Albrechtslund and Peter Lauritsen. 2013. Spaces of everyday surveillance: Unfolding an analytical concept of participation. *Geoforum* 49 (Oct. 2013), 310–316. https://doi.org/10.1016/j.geoforum.2013.04.016

[6] Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía González, Mariela Rajngewerc, Lautaro Martínez, Jorge Sánchez, Mauro Schilman, Guido Ivetta, Alexia Halvorsen, Amanda Mata Rojo, Matías Bordone, and Beatriz Busaniche. 2023. A methodology to characterize bias and harmful stereotypes in natural language processing in Latin America. http://arxiv.org/abs/2207.06591 arXiv:2207.06591 [cs].

[7] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* (Feb. 2023). https://doi.org/10.7759/cureus.35179

[8] Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2022. Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions. In *WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop*. 361 – 371. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152907136&partnerID=40&md5=fe18d9ed19b6dc3642108f3bd315942e Type: Conference paper.

[9] Dario Amodei, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. 2018. AI and compute. https://openai.com/research/ai-and-compute

[10] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. http://arxiv.org/abs/2307.03718 arXiv:2307.03718 [cs].

[11] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. https://doi.org/10.48550/arXiv.2007.03051 Issue: arXiv:2007.03051 arXiv: 2007.03051 [cs, eess, stat].

[12] E.-M. Anton, S. Devese, J. Miller, F. Ullstad, B.J. Ruck, H.J. Trodahl, and F. Natali. 2020. Superconducting computing memory using rare-earth nitrides. 92. Place: Australia.

[13] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J. Dobson, and James T. Teo. 2023. AI chatbots not yet ready for clinical use. *Frontiers in Digital Health* 5 (2023). https://doi.org/10.3389/fdgth.2023.1161098 Type: Article.

[14] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Foundational Models Defining a New Era in Vision: A Survey and Outlook. http://arxiv.org/abs/2307.13721 arXiv:2307.13721 [cs].

[15] Jeeyun (Sophia) Baik. 2020. Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA). *Telematics and Informatics* 52 (Sept. 2020), 101431. https://doi.org/10.1016/j.tele.2020.101431

[16] Francois Barnard, Marlize Van Sittert, and Sirisha Rambhatla. 2023. Self-Diagnosis and Large Language Models: A New Front for Medical Misinformation. http://arxiv.org/abs/2307.04910 arXiv:2307.04910 [cs].

[17] Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 2023. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. http://arxiv.org/abs/2206.08966 arXiv:2206.08966 [cs].

[18] Ulrich Beck. 1992. *Risk Society: Towards a New Modernity*. SAGE Publications. https://uk.sagepub.com/en-gb/eur/risk-society/book203184

[19] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, 610–623. https://doi.org/10.1145/3442188.3445922 Place: New York, NY, USA.

[20] W Lance Bennett and Steven Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication* 33, 2 (April 2018), 122–139. https://doi.org/10.1177/0267323118760317

[21] Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A Drop of Ink Makes a Million Think: The Spread of False Information in Large Language Models. http://arxiv.org/abs/2305.04812 arXiv:2305.04812 [cs].

[22] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* 5, 5 (May 2023), 277–280. https://doi.org/10.1038/s42254-023-00581-4

[23] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. http://arxiv.org/abs/2110.01963 arXiv:2110.01963 [cs].

[24] Su Lin Blodgett and Michael Madaio. 2021. Risks of AI Foundation Models in Education. http://arxiv.org/abs/2110.10024 arXiv:2110.10024 [cs].

[25] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 Issue: arXiv:2108.07258 arXiv: 2108.07258 [cs].

[26] Tore Bonsaksen, Mary Ruffolo, Janni Leung, Daicia Price, Hilde Thygesen, Mariyana Schoultz, and Amy Østertun Geirdal. 2021. Loneliness and Its Association With Social Media Use During the COVID-19 Outbreak. *Social Media + Society* 7, 3 (July 2021), 20563051211033821. https://doi.org/10.1177/20563051211033821

[27] Conrad Borchers, Dalia Sara Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M. Asano, and Hannah Rose Kirk. 2022. Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. http://arxiv.org/abs/2205.11374 arXiv:2205.11374 [cs].

[28] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. https://doi.org/10.48550/arXiv.2302.03494 arXiv:2302.03494 [cs].

[29] Leonie Bossert and Thilo Hagendorff. 2021. Animals and AI. The role of animals in AI research and application – An overview and ethical evaluation. *Technology in Society* 67, 101678 (2021). https://doi.org/10.1016/j.techsoc.2021.1

[30] Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *ACM International Conference Proceeding Series*. 379 – 388. https://doi.org/10.1145/3593013.3594005 Type: Conference paper.

[31] Rogers Brubaker. 2020. Digital hyperconnectivity and the self. *Theory and Society* 49 (Aug. 2020), 771–801. https://link.springer.com/article/10.1007/s11186-020-09405-1

[32] Thomas Andrew Bryer, Cristian Pliscoff, and Ashley Wilt Connors. 2020. *Promoting Civic Health Through University-Community Partnerships: Global Contexts and Experiences*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-19666-0

[33] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (Jan. 2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[34] Robert D. Bullard. 1993. The Threat of Environmental Racism. *Natural Resources & Environment* 7, 3 (1993), 23–56. https://www.jstor.org/stable/40923229

[35] Victoria Canning and Steve Tombs. 2021. *From Social Harm to Zemiology: A Critical Introduction* (first ed.). Routledge, London. https://doi.org/10.4324/9780429430497

[36] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. https://doi.org/10.48550/arXiv.2202.07646 Issue: arXiv:2202.07646 arXiv: 2202.07646 [cs].

[37] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. http://arxiv.org/abs/2306.09442 arXiv:2306.09442 [cs].

[38] Center for AI Safety. [n. d.]. Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk

[39] Anastasia Chan. 2023. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI and Ethics* 3, 1 (Feb. 2023), 53–64. https://doi.org/10.1007/s43681-022-00148-6

[40] Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. 2023. Reclaiming the Digital Commons: A Public Data Trust for Training Data. http://arxiv.org/abs/2303.09001 arXiv:2303.09001 [cs].

[41] P. V. Sai Charan, Hrushikesh Chunduri, P. Mohan Anand, and Sandeep K. Shukla. 2023. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads. http://arxiv.org/abs/2305.15336 arXiv:2305.15336 [cs].

[42] Yang Chen, Meena Andiappan, Tracy Jenkin, and Anton Ovchinnikov. 2023. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4380365

[43] Bartlomiej Chomanski. 2021. The Missing Ingredient in the Case for Regulating Big Tech. *Minds and Machines* 31, 2 (June 2021), 257–275. https://doi.org/10.1007/s11023-021-09562-x

[44] Miranda Christ, Sam Gunn, and Or Zamir. 2023. Undetectable Watermarks for Language Models. http://arxiv.org/abs/2306.09194 arXiv:2306.09194 [cs].

[45] Simon Coghlan and Christine Parker. 2023. Harm to Nonhuman Animals from AI: a Systematic Account and Framework. *Philosophy & Technology* 36, 2 (June 2023), 25. https://doi.org/10.1007/s13347-023-00627-6

[46] I. Glenn Cohen. 2023. What Should ChatGPT Mean for Bioethics? *American Journal of Bioethics* (2023). https://doi.org/10.1080/15265161.2023.2233357 Type: Article.

[47] Julie E. Cohen. 2019. *Between Truth and Power: The Legal Constructions of Informational Capitalism* (1 ed.). Oxford University PressNew York. https://doi.org/10.1093/oso/9780190246693.001.0001

[48] Ben Collier, Gemma Flynn, James Stewart, and Daniel Thomas. 2022. Influence government: Exploring practices, ethics, and power in the use of targeted advertising by the UK state. *Big Data & Society* 9, 1 (Jan. 2022), 20539517221078756. https://doi.org/10.1177/20539517221078756

[49] Mark Connor and Michael O'Neill. 2023. Large Language Models in Sport Science & Medicine: Opportunities, Risks and Considerations. http://arxiv.org/abs/2305.03851 arXiv:2305.03851 [cs].

[50] Jeremy W. Crampton. 2019. Digital Geographies. In *Digital Geographies*. SAGE Publications Ltd, 55 City Road, 281–290. https://doi.org/10.4135/9781529793536

[51] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t

[52] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241–1299. https://doi.org/10.2307/1229039

[53] Daniela America da Silva, Henrique Duarte Borges Louro, Gildarcio Sousa Goncalves, Johnny Cardoso Marques, Luiz Alberto Vieira Dias, Adilson Marques da Cunha, and Paulo Marcelo Tasinaffo. 2021. Could a conversational ai identify offensive language? *Information (Switzerland)* 12, 10 (2021). https://doi.org/10.3390/info12100418 Type: Article.

[54] Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022. Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference. In *Proceedings - International Conference on Computational Linguistics, COLING*, Vol. 29. 1442 – 1454. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165764100&partnerID=40&md5=23867432dc473bf00fedf132388e4412 Issue: 1 Type: Conference paper.

[55] S. de Jager. 2023. Semantic noise in the Winograd Schema Challenge of pronoun disambiguation. *Humanities and Social Sciences Communications* 10, 1 (April 2023), 1–10. https://doi.org/10.1057/s41599-023-01643-9 Number: 1 Publisher: Palgrave.

[56] Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American Language Bias in Natural Language Generation. http://arxiv.org/abs/2305.14291 arXiv:2305.14291 [cs].

[57] Ronald J. Deibert and Louis W. Pauly. 2019. Mutual entanglement and complex sovereignty in cyberspace. In *Data Politics*. Routledge.

[58] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. https://doi.org/10.48550/arXiv.2303.18190 Issue: arXiv:2303.18190 arXiv: 2303.18190 [cs].

[59] Erik Derner and Kristina Batistič. 2023. Beyond the Safeguards: Exploring the Security Risks of ChatGPT. http://arxiv.org/abs/2305.08005 Issue: arXiv:2305.08005 arXiv: 2305.08005 [cs].

[60] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization? http://arxiv.org/abs/2306.01248 arXiv:2306.01248 [cs].

[61] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. http://arxiv.org/abs/2304.05335 Issue: arXiv:2304.05335 arXiv: 2304.05335 [cs].

[62] Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y.-Lan Boureau, and Verena Rieser. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. http://arxiv.org/abs/2107.03451 arXiv:2107.03451 [cs].

[63] Georges Dionne. 2013. Risk Management: History, Definition, and Critique. *Risk Management and Insurance Review* 16, 2 (2013), 147–166. https://doi.org/10.1111/rmir.12016

[64] Andrés Domínguez Hernández, Kopo M. Ramokapane, Partha Das Chowdhury, Ola Michalec, Emily Johnstone, Emily Godwin, Alicia G. Cork, and Awais Rashid. 2023. Co-creating a Transdisciplinary Map of Technology-mediated Harms, Risks and Vulnerabilities: Challenges, Ambivalences and Opportunities. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 330:1–330:21. https://doi.org/10.1145/3610179

[65] Florin Eggmann, Roland Weiger, Nicola U. Zitzmann, and Markus B. Blatz. 2023. Implications of large language models such as ChatGPT for dental medicine. *JOURNAL OF ESTHETIC AND RESTORATIVE DENTISTRY* (April 2023). https://doi.org/10.1111/jerd.13046 Type: Review; Early Access.

[66] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. http://arxiv.org/abs/2303.10130 arXiv:2303.10130 [cs, econ, q-fin].

[67] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 153–162. https://doi.org/10.1145/2702123.2702556 Place: Seoul Republic of Korea.

[68] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Publishing Group.

[69] Mingyuan Fan, Cen Chen, Chengyu Wang, and Jun Huang. 2023. On the Trustworthiness Landscape of State-of-the-art Generative Models: A Comprehensive Survey. http://arxiv.org/abs/2307.16680 arXiv:2307.16680 [cs] version: 1.

[70] Mirko Farina and Andrea Lavazza. 2023. ChatGPT in society: emerging issues. *Frontiers in Artificial Intelligence* 6 (2023). https://doi.org/10.3389/frai.2023.1130913 Type: Article.

[71] Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or Foe? Exploring the Implications of Large Language Models on the Science System. http://arxiv.org/abs/2306.09928 arXiv:2306.09928 [cs].

[72] Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. http://arxiv.org/abs/2306.15087 Issue: arXiv:2306.15087 arXiv: 2306.15087 [cs].

[73] Gabriele Ferri and Inte Gloerich. 2023. Risk and Harm: Unpacking Ideologies in the AI Discourse. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, 1–6. https://doi.org/10.1145/3571884.3603751 Place: New York, NY, USA.

[74] Marion Fourcade and Jeffrey Gordon. 2020. Learning Like a State: Statecraft in the Digital Age. *Journal of Law and Political Economy* 1, 1 (2020). https://doi.org/10.5070/LP61150258

[75] Allie Funk, Adrian Shahbaz, and Kian Vesteinsson. 2023. Freedom of th eNEt 2023: The Repressive Power of Artificial Intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[76] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *ACM International Conference Proceeding Series*. 205 – 216. https://doi.org/10.1145/3593013.3593989

[77] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6, 1 (2023). https://doi.org/10.1038/s41746-023-00819-6 Type: Article.

[78] Timnit Gebru and Émile P. Torres. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* (April 2024). https://doi.org/10.5210/fm.v29i4.13636

[79] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. http://arxiv.org/abs/2009.11462 arXiv:2009.11462 [cs].

[80] Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. 2022. Risk of Bias in Chest X-ray Foundation Models. http://arxiv.org/abs/2209.02965 arXiv:2209.02965 [cs, eess].

[81] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. LLM Censorship: A Machine Learning Challenge or a Computer Security Problem? http://arxiv.org/abs/2307.10719 arXiv:2307.10719 [cs].

[82] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. http://arxiv.org/abs/2301.04246 arXiv:2301.04246 [cs].

[83] GOV.UK. 2023. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[84] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. http://arxiv.org/abs/2302.12173 Issue: arXiv:2302.12173 arXiv: 2302.12173 [cs].

[85] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 122–133. https://doi.org/10.1145/3461702.3462536 arXiv: 2006.03955 [cs].

[86] Thilo Hagendorff, Leonie Bossert, Tse Yip Fai, and Peter Singer. 2023. Speciesist bias in AI – How AI applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics* 3, 3 (Aug. 2023), 717–734. https://doi.org/10.1007/s43681-022-00199-9 arXiv: 2202.10848 [cs].

[87] Thilo Hagendorff and David Danks. 2023. Ethical and methodological challenges in building morally informed AI systems. *AI and Ethics* 3, 2 (May 2023), 553–566. https://doi.org/10.1007/s43681-022-00188-y

[88] Karen Hao. 2019. The computing power needed to train AI is now rising seven times faster than ever before. https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/

[89] Julian Hazell. 2023. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. http://arxiv.org/abs/2305.06972 Issue: arXiv:2305.06972 arXiv: 2305.06972 [cs].

[90] Will Douglas Heaven. 2022. Generative AI is changing everything. But what's left when the hype is gone? https://www.technologyreview.com/2022/12/16/1065005/generative-ai-revolution-art/

[91] Melissa Heikkilä. 2022. How AI-generated text is poisoning the internet. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/

[92] Gina Helfrich. 2024. The harms of terminology: why we should reject so-called "frontier AI". *AI and Ethics* (March 2024). https://doi.org/10.1007/s43681-024-00438-1

[93] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation Models and Fair Use. http://arxiv.org/abs/2303.15715 arXiv:2303.15715 [cs].

[94] Peter Henderson, Eric Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. 2023. Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models. http://arxiv.org/abs/2211.14946 arXiv:2211.14946 [cs].

[95] Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. http://arxiv.org/abs/2306.03950 arXiv:2306.03950 [cs].

[96] Mohammad Hosseini and Serge P. J. M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *RESEARCH INTEGRITY AND PEER REVIEW* 8, 1 (May 2023). https://doi.org/10.1186/s41073-023-00133-5 Type: Review.

[97] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. http://arxiv.org/abs/2303.11074 arXiv:2303.11074 [cs].

[98] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. http://arxiv.org/abs/2305.11391 Issue: arXiv:2305.11391 arXiv: 2305.11391 [cs].

[99] Yue Huang, Qihui Zhang, Philip S. Y, and Lichao Sun. 2023. TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models. http://arxiv.org/abs/2306.11507 Issue: arXiv:2306.11507 arXiv: 2306.11507 [cs].

[100] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots Enact Malignant Stereotypes. In *ACM International Conference Proceeding Series*. 743 – 756. https://doi.org/10.1145/3531146.3533138 Type: Conference paper.

[101] Committee on Artificial Intelligence (CAI). 2023. Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law. https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043

[102] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership Inference Attack Susceptibility of Clinical Language Models. http://arxiv.org/abs/2104.08305 arXiv:2104.08305 [cs].

[103] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544548.3581196 Type: Conference paper.

[104] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy*

of Sciences 120, 11 (March 2023), e2208839120. https://doi.org/10.1073/pnas.2208839120 Publisher: Proceedings of the National Academy of Sciences.

[105] Erik Jones and Jacob Steinhardt. 2022. Capturing Failures of Large Language Models via Human Cognitive Biases. https://doi.org/10.48550/arXiv.2202.12299 Issue: arXiv:2202.12299 arXiv: 2202.12299 [cs].

[106] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 10697–10707. https://proceedings.mlr.press/v162/kandpal22a.html

[107] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. 2023. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. http://arxiv.org/abs/2302.05733 arXiv:2302.05733 [cs].

[108] Yuhao Kang, Qianheng Zhang, and Robert Roth. 2023. The Ethics of AI-Generated Maps: A Study of DALLE 2 and Implications for Cartography. http://arxiv.org/abs/2304.10743 arXiv:2304.10743 [cs].

[109] Rabimba Karanjai. 2022. Targeted Phishing Campaigns using Large Scale Language Models. http://arxiv.org/abs/2301.00665 arXiv:2301.00665 [cs].

[110] Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology* 36, 2 (April 2023), 27. https://doi.org/10.1007/s13347-023-00606-x

[111] Sunder Ali Khowaja, Parus Khuwaja, and Kapal Dev. 2023. ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. http://arxiv.org/abs/2305.03123 arXiv:2305.03123 [cs].

[112] Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science* 380, 6651 (June 2023), 1222–1223. https://doi.org/10.1126/science.adi0248

[113] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. http://arxiv.org/abs/2303.05453 Issue: arXiv:2303.05453 arXiv: 2303.05453 [cs].

[114] Keith Kirkpatrick. 2023. The Carbon Footprint of Artificial Intelligence. *Commun. ACM* 66, 8 (Aug. 2023), 17–19. https://doi.org/10.1145/3603746

[115] Sarah Kreps and Doug Kriner. 2023. How AI Threatens Democracy. *Journal of Democracy* 34, 3 (Oct. 2023), 122–131. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

[116] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? *Science* 381, 6654 (July 2023), 138–138. https://doi.org/10.1126/science.adi8982 Publisher: American Association for the Advancement of Science.

[117] Patrick Yung Kang Lee, Ning F. Ma, Ig-Jae Kim, and Dongwook Yoon. 2023. Speculating on Risks of AI Clones to Selfhood and Relationships: Doppelganger-phobia, Identity Fragmentation, and Living Memories. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 91:1–91:28. https://doi.org/10.1145/3579524

[118] Florian Lehmann. 2023. Mixed-Initiative Interaction with Computational Generative Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, 1–6. https://doi.org/10.1145/3544549.3577061 Place: New York, NY, USA.

[119] David Leslie and Francesca Rossi. 2023. *ACM TechBrief: Generative Artificial Intelligence*. Association for Computing Machinery, New York, NY, USA.

[120] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step Jailbreaking Privacy Attacks on ChatGPT. http://arxiv.org/abs/2304.05197 arXiv:2304.05197 [cs].

[121] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. http://arxiv.org/abs/2304.03271 Issue: arXiv:2304.03271 arXiv: 2304.03271 [cs].

[122] Siyu Li, Jin Yang, and Kui Zhao. 2023. Are you in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks. http://arxiv.org/abs/2307.10337 arXiv:2307.10337 [cs].

[123] Xuan Li, Yonglin Tian, Peijun Ye, Haibin Duan, and Fei-Yue Wang. 2023. A Novel Scenarios Engineering Methodology for Foundation Models in Metaverse. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53, 4 (2023), 2148 – 2159. https://doi.org/10.1109/TSMC.2022.3228594 Type: Article.

[124] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. http://arxiv.org/abs/2306.01941 arXiv:2306.01941 [cs].

[125] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. 2023. Differentiate ChatGPT-generated and Human-written Medical Texts. http://arxiv.org/abs/2304.11567 arXiv:2304.11567 [cs].

[126] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, 1–6. https://doi.org/10.1145/3571884.3604316 Place: New York, NY, USA.

[127] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating Political Bias in Language Models Through Reinforced Calibration. http://arxiv.org/abs/2104.14795 arXiv:2104.14795 [cs].

[128] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt Injection attack against LLM-integrated Applications. http://arxiv.org/abs/2306.05499 arXiv:2306.05499 [cs].

[129] Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, and Jon Whittle. 2023. Towards Responsible AI in the Era of ChatGPT: A Reference Architecture for Designing Foundation Model-based AI Systems. http://arxiv.org/abs/2304.11090 arXiv:2304.11090 [cs].

[130] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. http://arxiv.org/abs/2211.02001 Issue: arXiv:2211.02001 arXiv: 2211.02001 [cs].

[131] Dieuwertje Luitse and Wiebke Denkena. 2021. The great transformer: Examining the role of large language models in the political economy of AI. *Big Data and Society* 8, 2 (2021). https://doi.org/10.1177/20539517211047734 Type: Article.

[132] Brady D. Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74, 5 (2023), 570 – 581. https://doi.org/10.1002/asi.24750 Type: Article.

[133] Deborah Lupton. 2013. *Risk* (second ed.). https://doi.org/10.4324/9780203070161

[134] Deborah Lupton. 2016. *The Quantified Self: A Sociology of Self-Tracking*. Polity Press, Cambridge.

[135] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. https://doi.org/10.48550/arXiv.2307.15810 Issue: arXiv:2307.15810 arXiv: 2307.15810 [cs].

[136] Donald MacKenzie and Judy Wajcman. 1999. *The social shaping of technology* (second ed.). Open University Press, Buckingham, UK.

[137] Hanna Maria Malik, Nea Lepinkäinen, Anne Alvesalo-Kuusi, and Mika Viljanen. 2022. Social harms in an algorithmic context. *Justice, Power and Resistance* 5, 3 (Dec. 2022), 193–207. https://doi.org/10.1332/OYUA8095

[138] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2022. Theorising Algorithmic Justice. *European Journal of Information Systems* 31, 3 (May 2022), 269–287. https://doi.org/10.1080/0960085X.2021.1934130

[139] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet. https://doi.org/10.48550/arXiv.2306.06130 Issue: arXiv:2306.06130 arXiv: 2306.06130 [cs].

[140] Puranjay Mattas. 2023. ChatGPT: A Study of AI Language Processing and its Implications. *International Journal of Research Publication and Reviews* 4 (Feb. 2023), 435–440. https://doi.org/10.55248/gengpi.2023.4218

[141] Cade Metz. 2023. How Could A.I. Destroy Humanity? *The New York Times* (June 2023). https://www.nytimes.com/2023/06/10/technology/ai-humanity.html

[142] Cade Metz. 2023. 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times* (May 2023). https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

[143] David Mhlanga. 2023. Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4354422

[144] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. 1816 – 1826. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149443328&partnerID=40&md5=0312417870119f60c7eb8b58c2a8f797 Type: Conference paper.

[145] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More Human than Human: Measuring ChatGPT Political Bias. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4372349

[146] Martin Májovský, Martin Černý, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. *Journal of Medical Internet Research* 25 (2023). https://doi.org/10.2196/46924 Type: Article.

[147] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. 2022. Defining organizational AI governance. *AI and Ethics* 2, 4 (Nov. 2022), 603–609. https://doi.org/10.1007/s43681-022-00143-x

[148] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. http://arxiv.org/abs/2004.09456 Issue: arXiv:2004.09456 arXiv: 2004.09456 [cs].

[149] Luca Nannini. 2023. Voluminous yet Vacuous? Semantic Capital in an Age of Large Language Models. http://arxiv.org/abs/2306.01773 arXiv:2306.01773 [cs].

[150] Subash Neupane, Ivan A. Fernandez, Sudip Mittal, and Shahram Rahimi. 2023. Impacts and Risk of Generative AI Technology on Cyber Defense. http://arxiv.org/abs/2306.13033 Issue: arXiv:2306.13033 arXiv: 2306.13033 [cs].

[151] NIST. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Technical Report NIST AI 100-1. National Institute of Standards and Technology

(U.S.), Gaithersburg, MD. NIST AI 100–1 pages. https://doi.org/10.6028/NIST.AI.100-1

[152] Rob Nixon. 2013. *Slow Violence and the Environmentalism of the Poor.* https://www.hup.harvard.edu/books/9780674072343

[153] Debora Nozza, Federcio Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

[154] OECD. 2022. *Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint.* OECD Digital Economy Papers 341. https://doi.org/10.1787/7babf571-en

[155] OECD. 2023. *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI.* Technical Report. https://doi.org/10.1787/bf3c0c60-en

[156] OECD. 2023. *Initial policy considerations for generative artificial intelligence.* OECD Artificial Intelligence Papers 1. https://doi.org/10.1787/fae2d1e6-en Series: OECD Artificial Intelligence Papers Volume: 1.

[157] Kieron O'Hara, Wendy Hall, Vinton Cerf, Kieron O'Hara, Wendy Hall, and Vinton Cerf. 2021. *Four Internets: Data, Geopolitics, and the Governance of Cyberspace.* Oxford University Press, Oxford, New York.

[158] Michael O'Neill and Mark Connor. 2023. Amplifying Limitations, Harms and Risks of Large Language Models. http://arxiv.org/abs/2307.04821 Issue: arXiv:2307.04821 arXiv: 2307.04821 [cs].

[159] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *ACM International Conference Proceeding Series*. 1246 – 1266. https://doi.org/10.1145/3593013.3594078

[160] Oscar Oviedo-Trespalacios, Amy E. Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J. E. Rod., Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, Steffen Steinert, Ashleigh Filtness, and Genserik Reniers. 2023. The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice. https://doi.org/10.2139/ssrn.4346827

[161] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. https://doi.org/10.48550/arXiv.2305.13661 Issue: arXiv:2305.13661 arXiv: 2305.13661 [cs].

[162] Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine: The Struggle for Effective Text Anonymisation in the Age of Large Language Models. http://arxiv.org/abs/2303.12429 arXiv:2303.12429 [cs].

[163] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. http://arxiv.org/abs/2104.10350 arXiv:2104.10350 [cs].

[164] Fábio Perez and Ian Ribeiro. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. http://arxiv.org/abs/2211.09527 Issue: arXiv:2211.09527 arXiv: 2211.09527 [cs].

[165] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. 1291 – 1292. https://doi.org/10.1145/3539597.3575792 Type: Conference paper.

[166] Andres Piñeiro-Martin, Carmen Garcia-Mateo, Laura Docio-Fernandez, and Maria del Carmen Lopez-Perez. 2023. Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models. *ELECTRONICS* 12, 14 (July 2023). https://doi.org/10.3390/electronics12143170 Type: Article.

[167] Richard Plant, Valerio Giuffrida, and Dimitra Gkatzia. 2022. You Are What You Write: Preserving Privacy in the Era of Large Language Models. http://arxiv.org/abs/2204.09391 arXiv:2204.09391 [cs].

[168] Sebastian Porsdam Mann, Brian D. Earp, Nikolaj Møller, Suren Vynn, and Julian Savulescu. 2023. AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle. *American Journal of Bioethics* (2023). https://doi.org/10.1080/15265161.2023.2233356 Type: Article.

[169] Van Rensselaer Potter. 1999. Fragmented Ethics and "Bridge Bioethics". *The Hastings Center Report* 29, 1 (1999), 38–40. https://doi.org/10.2307/3528538

[170] Michael Power. 2009. The risk management of nothing. *Accounting, Organizations and Society* 34, 6 (Aug. 2009), 849–855. https://doi.org/10.1016/j.aos.2009.06.001

[171] Junaid Qadir. 2022. Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. https://doi.org/10.36227/techrxiv.21789434.v1

[172] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual Adversarial Examples Jailbreak Large Language Models. http://arxiv.org/abs/2306.13213 arXiv:2306.13213 [cs].

[173] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 959–972. https://doi.org/10.1145/3531146.3533158 Place: Seoul Republic of Korea.

[174] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. https://doi.org/10.48550/arXiv.2305.14965 Issue: arXiv:2305.14965 arXiv: 2305.14965 [cs].

[175] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140426523&partnerID=40&md5=97e0a22d17d6060168a661063e8587ed Type: Conference paper.

[176] Nathaniel A. Raymond. 2017. Beyond "Do No Harm" and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society's Use of Data. In *Group Privacy: New Challenges of Data Technologies*, Linnet Taylor, Luciano Floridi, and Bart van der Sloot (Eds.). Springer International Publishing, Cham, 67–82. https://doi.org/10.1007/978-3-319-46608-8_4

[177] Ludovico Rella. 2023. Close to the metal: Towards a material political economy of the epistemology of computation. *Social Studies of Science* 0, 0 (2023). https://doi.org/10.1177/030631272311850

[178] Matthias C. Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A. Gould, and Uli Sauerland. 2023. Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology* 57, 9 (March 2023), 3464–3466. https://doi.org/10.1021/acs.est.3c01106

[179] Scott Robbins and Aimee van Wynsberghe. 2022. Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. *Sustainability* 14, 8 (Jan. 2022), 4829. https://doi.org/10.3390/su14084829

[180] Loius Rosenberg. 2023. The Metaverse and Conversational AI as a Threat Vector for Targeted Influence. IEEE. https://doi.org/10.1109/CCWC57344.2023.10099167 Place: Las Vegas, NV, USA.

[181] David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences* 12, 3 (March 2023), 148. https://doi.org/10.3390/socsci12030148 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[182] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching* 6, 1 (Jan. 2023). https://doi.org/10.37074/jalt.2023.6.1.9

[183] Ebony L. Ruhland, Lauren Johnson, Janet Moore, Cinnamon Pelly, Simone Bess, and Jacinda K. Dariotis. 2023. Positionality, intersectionality, power dynamics in community participatory research to define public safety in Black communities. *Journal of Community Psychology* 51, 7 (2023), 2845–2860. https://doi.org/10.1002/jcop.23046

[184] Téo Sanchez. 2023. Examining the Text-to-Image Community of Practice: Why and How do People Prompt Generative AIs?. In *Proceedings of the 15th Conference on Creativity and Cognition (C&C '23)*. Association for Computing Machinery, 43–61. https://doi.org/10.1145/3591196.3593051 Place: New York, NY, USA.

[185] Jonas B. Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. http://arxiv.org/abs/2306.13952 arXiv:2306.13952 [cs].

[186] Marc Schuilenburg. 2017. *The Securitization of Society: Crime, Risk, and Social Order.* NYU Press.

[187] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. http://arxiv.org/abs/1907.10597 Issue: arXiv:1907.10597 arXiv: 1907.10597 [cs, stat].

[188] Glorin Sebastian. 2023. Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk? - An Exploratory Study. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4363843

[189] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. http://arxiv.org/abs/2212.08061 arXiv:2212.08061 [cs].

[190] Murray Shanahan. 2023. Talking About Large Language Models. http://arxiv.org/abs/2212.03551 arXiv:2212.03551 [cs].

[191] Hanyin Shao, Jie Huang, Shen Zheng, and Kevin Chen-Chuan Chang. 2023. Quantifying Association Capabilities of Large Language Models and Its Implications on Privacy Leakage. http://arxiv.org/abs/2305.12707 arXiv:2305.12707 [cs].

[192] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, 723–741. https://doi.org/10.1145/3600211.3604673 Place: New York, NY, USA.

[193] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim,

Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. 2023. Model evaluation for extreme risks. https://doi.org/10.48550/arXiv.2305.15324 arXiv:2305.15324 [cs].

[194] Katharina Simbeck. 2022. FAccT-Check on AI regulation: Systematic Evaluation of AI Regulation on the Example of the Legislation on the Use of AI in the Public Sector in the German Federal State of Schleswig-Holstein. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, 89–96. https://doi.org/10.1145/3531146.3533076 Place: New York, NY, USA.

[195] Gabriel Simmons. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. https://doi.org/10.48550/arXiv.2209.12106 Issue: arXiv:2209.12106 arXiv: 2209.12106 [cs].

[196] Nathalie A. Smuha. 2021. Beyond the individual: governing AI's societal harm. *Internet Policy Review* 10, 3 (Sept. 2021). https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm

[197] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? http://arxiv.org/abs/2306.03809 arXiv:2306.03809 [cs].

[198] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. https://doi.org/10.48550/arXiv.1908.09203 Issue: arXiv:1908.09203 arXiv: 1908.09203 [cs].

[199] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. http://arxiv.org/abs/2306.05949 arXiv:2306.05949 [cs].

[200] Bernd Carsten Stahl, Josephina Antoniou, Mark Ryan, Kevin Macnish, and Tilimbe Jiya. 2022. Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY* 37, 1 (March 2022), 23–37. https://doi.org/10.1007/s00146-021-01148-6

[201] Susan Leigh Star. 1999. The Ethnography of Infrastructure. *American Behavioral Scientist* 43, 3 (Nov. 1999), 377–391. https://doi.org/10.1177/00027649921955326

[202] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* 48, 2 (May 2018). https://doi.org/10.1177/0306312718772094

[203] Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 3524 – 3542. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140424088&partnerID=40&md5=c8c22ce85f123e869740315aa1a8f53d

[204] Mariya Stoilova, Sonia Livingstone, and Rana Khazbak. 2021. Investigating Risks and Opportunities for Children in a Digital World: A rapid review of the evidence on children's internet use and outcomes. (2021). https://www.end-violence.org/sites/default/files/paragraphs/download/Investigating-Risks-and-Opportunities-for-Children-in-a-Digital-World.pdf

[205] Christian Stoll, Lena Klaaßen, and Ulrich Gallersdörfer. 2019. The Carbon Footprint of Bitcoin. *Joule* 3, 7 (July 2019), 1647–1661. https://doi.org/10.1016/j.joule.2019.05.012

[206] Anna Strasser. 2023. On pitfalls (and advantages) of sophisticated large language models. http://arxiv.org/abs/2303.17511 arXiv:2303.17511 [cs].

[207] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. http://arxiv.org/abs/1906.02243 Issue: arXiv:1906.02243 arXiv: 1906.02243 [cs].

[208] Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety Assessment of Chinese Large Language Models. http://arxiv.org/abs/2304.10436 arXiv:2304.10436 [cs].

[209] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, virtual+Dublin, 26–41. https://doi.org/10.18653/v1/2022.bigscience-1.3

[210] Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption. http://arxiv.org/abs/2304.01358 arXiv:2304.01358 [cs].

[211] Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. http://arxiv.org/abs/1911.01485 Issue: arXiv:1911.01485 arXiv: 1911.01485 [cs, stat].

[212] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The Science of Detecting LLM-Generated Texts. http://arxiv.org/abs/2303.07205 arXiv:2303.07205 [cs].

[213] Iddo Tavory and Stefan Timmermans. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Univeristy of Chicago Press. https://doi.org/10.7208/chicago/9780226180458.001.0001

[214] Vishesh Thakur. 2023. Unveiling Gender Bias in Terms of Profession Across LLMs: Analyzing and Addressing Sociological Implications. http://arxiv.org/abs/2307.09162 arXiv:2307.09162 [cs].

[215] Surendrabikram Thapa and Surabhi Adhikari. 2023. ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. *ANNALS OF BIOMEDICAL ENGINEERING* (June 2023). https://doi.org/10.1007/s10439-023-03284-0 Type: Article; Early Access.

[216] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544549.3583177 Type: Conference paper.

[217] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. http://arxiv.org/abs/2306.10070 arXiv:2306.10070 [cs, q-bio].

[218] Isaac Triguero, Daniel Molina, Javier Poyatos, Javier Del Ser, and Francisco Herrera. 2023. General Purpose Artificial Intelligence Systems (GPAIS): Properties, Definition, Taxonomy, Open Challenges and Implications. http://arxiv.org/abs/2307.14283 arXiv:2307.14283 [cs].

[219] Félix Tréguer. 2019. *Seeing like Big Tech: security assemblages, technology, and the future of state bureaucracy*.

[220] Yariv Tsfati, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren. 2020. Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Annals of the International Communication Association* 44, 2 (April 2020), 157–173. https://doi.org/10.1080/23808985.2020.1759443

[221] UNESCO. 2023. Taking a human rights-based approach to artificial intelligence and the rule of law: UNESCO at the Athens Roundtable 2021. https://www.unesco.org/en/articles/taking-human-rights-based-approach-artificial-intelligence-and-rule-law-unesco-athens-roundtable

[222] United Nations. [n. d.]. Goal 13: Take urgent action to combat climate change and its impacts. https://www.un.org/sustainabledevelopment/climate-change/

[223] Aleksandra Urman and Mykola Makhortykh. 2023. The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. (2023).

[224] Cuma Uz and Ebru Umay. 2023. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? *International Journal of Rheumatic Diseases* 26, 7 (2023), 1343 – 1349. https://doi.org/10.1111/1756-185X.14749 Type: Article.

[225] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (Jan. 2020), 2056305120903408. https://doi.org/10.1177/2056305120903408

[226] Shannon Vallor and Ewa Luger. 2023. A shrinking path to safety: how a narrowly technical approach to align AI with the public good could fail. http://efi.ed.ac.uk/a-shrinking-path-to-safety-how-a-narrowly-technical-approach-to-align-ai-with-the-public-good-could-fail/

[227] Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2023. Undesirable biases in NLP: Averting a crisis of measurement. http://arxiv.org/abs/2211.13709 Issue: arXiv:2211.13709 arXiv: 2211.13709 [cs].

[228] Aniket Vashishtha, S. Sai Krishna Prasad, Payal Bajaj, Vishrav Chaudhary, Kate Cook, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2023. Performance and Risk Trade-offs for Multi-word Text Prediction at Scale. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*. 2181 – 2197. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159852354&partnerID=40&md5=9cfbc817cef9c008340b930e1e236096 Type: Conference paper.

[229] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[230] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. http://arxiv.org/abs/2302.02463 arXiv:2302.02463 [cs].

[231] Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, and Srijan Kumar. 2022. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. http://arxiv.org/abs/2205.09744 arXiv:2205.09744 [cs].

[232] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. http://arxiv.org/abs/2306.07899 arXiv:2306.07899 [cs].

[233] Luis Vila-Henninger, Claire Dupuy, Virginie Van Ingelgom, Mauro Caprioli, Ferdinand Teuber, Damien Pennetreau, Margherita Bussi, and Cal Le Gall. 2022. Abductive Coding: Theory Building and Qualitative (Re)Analysis. *Sociological Methods & Research* (Feb. 2022), 00491241211067508. https://doi.org/10.1177/00491241211067508

[234] Christian von Scheve and Markus Lange. 2023. Risk entanglement and the social relationality of risk. *Humanities and Social Sciences Communications* 10, 1 (April 2023), 1–10. https://doi.org/10.1057/s41599-023-01668-0 Publisher: Palgrave.

[235] Lauren Walker. 2023. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times* (March 2023). https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt

[236] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. https://doi.org/10.48550/arXiv.2306.11698 Issue: arXiv:2306.11698 arXiv: 2306.11698 [cs].

[237] Hofit Wasserman Rozen, Niva Elkin-Koren, and Ran Gilad-Bachrach. 2023. The Case Against Explainability. http://arxiv.org/abs/2305.12167 arXiv:2305.12167 [cs].

[238] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. http://arxiv.org/abs/2112.04359 arXiv:2112.04359 [cs].

[239] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, 214–229. https://doi.org/10.1145/3531146.3533088 Place: New York, NY, USA.

[240] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. http://arxiv.org/abs/2109.07445 arXiv:2109.07445 [cs].

[241] Sarah Jane O White and James P Shine. 2016. Exposure Potential and Health Impacts of Indium and Gallium, Metals Critical to Emerging Electronics and Energy Technologies. *Current environmental health reports* 3, 4 (Dec. 2016), 459–467. https://doi.org/10.1007/s40572-016-0118-8

[242] Clare Williams. 2022. Framing the Future: The Foundation Series, Foundation Models and Framing AI. *LAW TECHNOLOGY AND HUMANS* 4, 2 (2022), 109–123. https://doi.org/10.5204/lthj.2452 Type: Article.

[243] Robin Williams and David Edge. 1996. The social shaping of technology. *Research Policy* 25, 6 (Sept. 1996), 865–899. https://doi.org/10.1016/0048-7333(96)00885-2

[244] Bernd W. Wirtz, Jan C. Weyerer, and Ines Kehl. 2022. Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly* 39, 4 (Oct. 2022), 101685. https://doi.org/10.1016/j.giq.2022.101685

[245] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 1–10. https://doi.org/10.1145/2601248.2601268 Place: London England United Kingdom.

[246] Claes Wohlin, Marcos Kalinowski, Katia Romero Felizardo, and Emilia Mendes. 2022. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology* 147 (July 2022), 106908. https://doi.org/10.1016/j.infsof.2022.106908

[247] Xiaodong Wu, Ran Duan, and Jianbing Ni. 2023. Unveiling Security, Privacy, and Ethical Concerns of ChatGPT. http://arxiv.org/abs/2307.14192 arXiv:2307.14192 [cs].

[248] Malwina Anna Wójcik. 2022. Foundation Models in Healthcare: Opportunities, Biases and Regulatory Prospects in Europe. In *ELECTRONIC GOVERNMENT AND THE INFORMATION SYSTEMS PERSPECTIVE, EGOVIS 2022 (Lecture Notes in Computer Science, Vol. 13429)*, A Ko, E Francesconi, G Kotsis, AM Tjoa, and I Khalil (Eds.). 32–46. https://doi.org/10.1007/978-3-031-12673-4_3 Type: Proceedings Paper.

[249] Chloe Xiang. 2023. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. *VICE* (March 2023). https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

[250] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2023. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. http://arxiv.org/abs/2303.13379 arXiv:2303.13379 [cs].

[251] Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. 2023. Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. http://arxiv.org/abs/2305.10235 arXiv:2305.10235 [cs].

[252] Ali Zarifhonarvar. 2023. Economics of ChatGPT: A Labor Market View on the Occupational Impact of Artificial Intelligence. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4350925

[253] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2023. Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions. http://arxiv.org/abs/2307.03941 arXiv:2307.03941 [cs].

[254] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. http://arxiv.org/abs/2303.18223 Issue: arXiv:2303.18223 arXiv: 2303.18223 [cs].

[255] Jianlong Zhou, Heimo Müller, Andreas Holzinger, and Fang Chen. 2023. Ethical ChatGPT: Concerns, Challenges, and Commandments. http://arxiv.org/abs/2305.10646 arXiv:2305.10646 [cs].

[256] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3581318 event-place: Hamburg, Germany.

[257] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. http://arxiv.org/abs/2301.12867 arXiv:2301.12867 [cs].

[258] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. http://arxiv.org/abs/2301.12867 Issue: arXiv:2301.12867 arXiv: 2301.12867 [cs].

A APPENDIX

A.1 Search strategy and objectives

The primary goal of our search strategy was to identify relevant literature on risks and harms of foundation models and ensure a comprehensive understanding of their socio-political and environmental dimensions. To achieve this objective, we conducted searches across five academic electronic databases. The selected databases include arXiv, ACM Digital Library, IEEE, Scopus, and Web of Science. These databases were chosen based on their relevance in the field, their inclusion of a broader range of sources such as conference proceedings and preprints, and access to the full text of the articles.

To develop the search strategy, we chose a handful of keywords based on their relevance to the research question and maximize the scope of the search: "foundation model", "large language model", "llm", "general purpose artificial intelligence", "GPAI", "risk", "harm", "ethic". In addition, we used Boolean operators to refine search queries (see Table 2 for example). In some cases, we refined when the result set had a high percentage of irrelevant returns: AND TI "artificial intelligence" OR TI "AI" OR AB "artificial intelligence" OR AB "AI".

Searches were conducted on title and abstract fields. A sample set was piloted, before a full search was conducted across the selected databases. Inclusion criteria covered English language journal articles, including conference papers and pre-prints, but excluding grey literature, monographs, commentaries, correspondences, and opinion pieces.

The results were added to a Zotero library, which also included our snowball sample, and the merge of these two libraries was exported into a spreadsheet for refinement and coding. We then removed duplicates and non-relevant papers, reaching 167 papers in our sample. The use of an abductive approach supported establishing clearer links with the research objectives and developing a codified mapping of key relationships found in the literature results. To code the papers, firstly, we extracted applicable keyword information from the abstract of each primary study as part of an initial coding process. The resulting keywords served largely as summative and process attributes of a risk or type of harm relevant to foundation models or LLMs. Three researchers subsequently clustered these attributes (Level 2 – attribute codes) and assigned them into Level 1 or parent categories. For example, to attribute code *copyright and intellectual property violations* we assigned the parent category *legal and regulatory violations*. For the attribute codes *difficult to ensure explainability* and *reliance on proprietary software* we assigned the parent category *lock-in and opacity risks* (See Table 3 for full list of parent and attribute codes). The codes were subsequently refined by all co-authors.

Table 2: Example of broad search (with refining keywords):

Broad search (*)	Refining #2
TI "foundation model*" OR AB "foundation model*"	AND TI "Risk*" OR TI "Harm*" OR TI "ethic*"OR AB "Risk*" OR AB "Harm*" OR AB "ethic*"
TI "Large Language model*" OR TI "LLM*" OR AB "Large Language model*" OR AB "LLM*"	AND TI "Risk*" OR TI "Harm*" OR TI "ethic*" OR AB "Risk*" OR AB "Harm*" OR AB "ethic*"
TI "general purpose artificial intelligence" OR TI "general-purpose artificial intelligence" OR TI "GPAI" OR AB "general purpose artificial intelligence" OR AB "general-purpose artificial intelligence" OR AB "GPAI"	AND TI "Risk*" OR TI "Harm*" OR TI "ethic*" OR AB "Risk*" OR AB "Harm*" OR AB "ethic*"

Table 3: Parent and attribute codes.

Parent Code	Attribute Code
Bias and societal prejudices	Amplifying and perpetuating stereotypes and societal biases
	Behavioral biases
	Political biases
Misinformation, disinformation and propaganda	Abusive interactive experiences
	Manipulation
	Misinformation spread
Unreliable performance	Accuracy/inaccuracy (outputs)
	Harmful or toxic outputs
	Language performance gap
	Poor performance due to excessive training with synthetic data
	Unpredictability of behaviour pre- and post-deployment
	Untruthfulness of outputs
	Disparate performance
Cybersecurity risks and harms	Cyber-attacks payload
	Data leakage
	Data poisoning attacks
	Fraudulent services
	Goal hijacking and prompt leaking
	Prompt injection attacks
	Spear phishing
	Jailbreaking
	Malicious actors
	Impersonation attack
Privacy risks and harms	Privacy risks and harms
Systemic social and economic risks and harms	Erosion of semantic capital
	Misleading narratives about AI
	Narrowing of the market/Monopolisation
	Outcome homogenization
	Perpetuation of inequalities
	Effects on labor market
	Widening of digital divide
	Concentration of authoritative power
	Invisibilization and poor working condition of data and content moderation labor
Legal and regulatory violations	Copyright and intellectual property violations
	Data protection violations
	Consumer protection laws violations
Environmental effects and ecological disruption	Environmental effects and ecological disruption
Misuse	Misuse
	Biological misuse
	Dual use
	Illegitimate surveillance
	Creation of violent or harmful content
Lock-in and opacity risks	Challenges to benchmarking
	Difficult to ensure explainability
	Lack of replicability and transparency
	Low technological readiness
	Reliance on proprietary software
Overdependency in human-computer interaction	Overdependency in human-computer interaction
Data risks and harms	Data extractivism
	Data quality
	Datasets containing toxic data
	Degradation of the digital commons
Value misalignment	Value misalignment
Extreme or catastrophic risks and harms	Extreme risks
	Catastrophic risks

A.2 Results of the coding and codes description

Table 4: Papers per risks and harms parent and attribute codes and codes description.

Risks and Harms Codes (and Papers)	Code Description	Count
Bias and societal prejudices Amplifying and perpetuating stereotypes and societal biases: [1, 2, 6, 10, 13, 14, 19, 23, 25, 27, 30, 40, 49, 53–56, 69, 70, 72, 76, 80, 85, 95, 96, 98–100, 103, 111–113, 124, 140, 143, 148, 153, 158, 159, 166, 171, 175, 188, 189, 192, 195, 198, 199, 203, 208–211, 214, 216, 227, 230, 236, 238, 239, 247, 248, 254, 258] Behavioral biases: [42] Political biases: [127, 145, 181, 223]	The perpetuation and amplification of hegemonic views, harmful stereotypes, societal and behavioral biases.	72
Misinformation, disinformation and propaganda Abusive interactive experiences: [29, 180, 192] Disinformation: [10, 16, 21, 25, 44, 71, 112, 123, 125, 140, 146, 150, 161, 192, 198, 199, 238, 239] Manipulation: [19, 40, 113, 192, 193, 238, 239] Hallucinated information: [49, 60, 108, 124, 146, 217] Misinformation spread: [10, 21, 39, 44, 70, 71, 77, 82, 98, 104, 111–113, 123–125, 140, 150, 166, 171, 178, 188, 192, 199, 206, 208, 212, 215, 238, 239, 255, 256, 258] .	Unintentional or intentional efforts to disseminate false or misleading information. Related risks and harms include disinformation, misinformation spread, extreme manipulation, and abusive interactive experiences.	45
Cybersecurity risks and harms Cyber-attacks payload: [41, 238, 239] Data leakage: [19, 59, 69, 98, 102, 124, 158, 166, 167, 188, 191, 236, 238, 239, 255, 258] Data poisoning attacks: [69, 98, 120, 247] Fraudulent services: [59, 70, 113, 168, 188, 206, 238, 239] Goal hijacking and prompt leaking: [164] Prompt injection attacks: [84, 128, 247, 258] Spear phishing: [10, 59, 65, 70, 89, 109, 150, 188, 199] Jailbreaking: [174, 236] Malicious actors: [14, 25, 65, 70, 104, 107, 238, 239, 247] Impersonation attack: [166, 199, 255]	Unauthorized access and attacks to the foundation model that exploit vulnerabilities or compromise its integrity. Examples of cybersecurity risks and harms include cyber-attacks payload, data leakage, data poisoning attacks, data pollution, fraudulent services, goal hijacking and prompt leaking, offensive cyber capabilities, prompt injection attacks, spear phishing, jailbreaking, and other attacks by malicious actors.	37

Risks and Harms Codes (and Papers)	Code Description	Count
Unreliable performance Accuracy/inaccuracy (outputs): [7, 13, 14, 28, 37, 70, 71, 98, 108, 124, 158, 166, 182, 192, 206, 224, 251, 258] Harmful or toxic outputs: [10, 19, 24, 28, 37, 53, 61, 62, 79, 99, 122, 124, 135, 172, 175, 189, 192, 199, 206, 208, 218, 228, 238–240, 257, 258] Language performance gap: [8, 231, 238, 239, 258] Poor performance due to excessive training with synthetic data: [232] Unpredictability of behaviour pre- and post- deployment: [10] Untruthfulness of outputs: [28, 160, 175] Disparate performance: [25, 192, 199]	Inconsistent or undesirable performance exhibited by foundation models. This includes inaccurate or non-factual outputs, harmful or toxic outputs, language performance gap, performance disparities at group levels, and unpredictability of behavior pre- and post-deployment.	44
Privacy risks and harms [19, 65, 71, 98, 102, 106, 111, 113, 120, 124, 129, 144, 158, 162, 165, 166, 168, 192, 199, 206, 217, 236, 238, 239, 247, 250, 253, 255]	Adverse consequences associated with the collection, storage, processing, and use of data beyond the data pipeline that arise from the increasing capabilities of foundation models and their potential to inadvertently expose private or sensitive data.	29
Systemic social and economic risks and harms Erosion of semantic capital: [149] Misleading narratives about AI: [73, 149, 190] Narrowing of the market/Monopolisation: [25, 131] Outcome homogenization: [25, 113, 168, 216] Perpetuation of inequalities: [25, 71, 113, 168, 178, 192, 199, 238, 239, 242] Effects on labor market: [66, 70, 111, 113, 124, 132, 140, 171, 199, 238, 239, 252] Widening of digital divide: [111, 192] Concentration of authoritative power: [111, 199] Invisibilization and poor working condition of data and content moderation labor: [199]	Risks and harms that arise from the widespread adoption and impact of foundation models on societal and economic structures. These range from the erosion of semantic capital, outcome homogenization, and misleading narratives about AI, to the widening of digital divide, the narrowing of the market or monopolization, effects on labor market, and the perpetuation of inequalities.	24
Legal and regulatory violations Copyright and intellectual property violations: [3, 25, 70, 71, 93, 97, 111, 117, 132, 158, 166, 168, 188, 199, 206] Data protection violations: [3, 25, 106, 111, 113, 158, 162, 165, 166, 168, 188, 199, 206, 217, 253] Consumer protection laws violations: [188]	Breaches of laws and regulations, including those designed to safeguard individuals' privacy and secure handling of their personal data, intellectual property, consumer protection, and cybersecurity.	21

Risks and Harms Codes (and Papers)	Code Description	Count
Environmental effects and ecological disruption [11, 19, 25, 86, 98, 111, 113, 114, 121, 124, 130, 158, 163, 178, 192, 199, 218, 239]	Adverse and negative impacts of foundation models on ecology and animals.	19
Misuse Misuse: [10] Biological misuse: [46, 185] Dual use: [10, 81, 94, 107, 197, 238, 239] Illegitime surveillance: [10, 25, 113, 238, 239] Creation of violent or harmful content: [10, 19, 25, 113, 192, 199]	The misuse of foundation models is used here to describe specific contexts in which foundation models or technologies based on foundation models may be inappropriately used and lead to negative consequences. This includes uses that can have both civilian and military applications, or contribute to the handling, manipulation, or application of biological materials, organisms, or technologies.	19
Lock-in and opacity risks Challenges to benchmarking: [14] Difficult to ensure explainability: [166, 216, 218, 237] Lack of replicability and transparency: [70, 96, 108, 111, 124, 129, 158, 166, 250, 255] Low technological readiness: [250] Reliance on proprietary software: [124, 126]	Risks and harms associated with foundation models that are difficult to understand, replicate, or modify. These risks and harms arise from issues including reliance on proprietary software, challenges to benchmarking, lack of transparency, difficulty to ensure explainaibility, and low technological readiness.	15
Overdependency in human-computer interaction [49, 113, 124, 135, 143, 165, 166, 199, 216, 238, 239]	Overreliance of humans interacting with computer systems, interfaces, and technologies built on foundation models for various aspects of their lives and potentially negatively impacting their well-being, safety, decision-making abilities, or interpersonal relationships.	11
Data risks and harms Data extractivism: [97] Data quality: [97, 106] Datasets containing toxic data: [22, 69, 158, 166, 236] Degradation of the digital commons: [39, 97]	A range of risks and harms associated with the collection, storage, processing, and use of data within the data pipeline. These risks extend beyond the cybersecurity risks and harms described above and includes broader issues of ethical, social, and economic implications of data stewardship and management. They include issues of poor data quality, datasets containing toxic data, data extractivism, and the degradation of the digital commons.	8
Value misalignment [97, 110]	Misalignment of the output content of conversational agents with the norms and values of the human interacting with such agent.	2
Extreme or catastrophic risks and harms Extreme risks: [193] Catastrophic risks: [17]	Speculative far-reaching or irreversible adverse impacts of foundation models at societal scale that extend beyond immediate impacts.	2