ABSTRACT

Responsible innovation on large-scale Language Models (LMs) requires foresight into and in-depth understanding of the risks these models may pose. This paper develops a comprehensive taxonomy of ethical and social risks associated with LMs. We identify twenty-one risks, drawing on expertise and literature from computer science, linguistics, and the social sciences. We situate these risks in our taxonomy of six risk areas: I. Discrimination, Hate speech and Exclusion, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, and VI. Environmental and Socioeconomic harms. For risks that have already been observed in LMs, the causal mechanism leading to harm, evidence of the risk, and approaches to risk mitigation are discussed. We further describe and analyse risks that have not yet been observed but are anticipated based on assessments of other language technologies, and situate these in the same taxonomy. We underscore that it is the responsibility of organizations to engage with the mitigations we discuss throughout the paper. We close by highlighting challenges and directions for further research on risk evaluation and mitigation with the goal of ensuring that language models are developed responsibly.

CCS CONCEPTS

• General and reference → Evaluation, Surveys and overviews; • Human-centered computing → HCI theory, concepts and models; • Social and professional topics → Computing / technology policy.

KEYWORDS

language models, responsible innovation, technology risks, responsible AI, risk assessment

ACM Reference Format:

1 INTRODUCTION

Language Model (LM) research is growing in scale and achieving research breakthroughs [30, 46, 147, 148, 170]. Several Artificial Intelligence (AI) research labs are pursuing LM research, spurred by the promise these models hold for a wide range of beneficial real-world applications. Large-scale LMs can be adapted to a wide
range of downstream tasks, creating the potential to affect many aspects of life [24]. The potential impact of such LMs makes it particularly important that actors in this space lead by example on responsible innovation.

Responsible innovation entails that in addition to developing a given technology, innovators anticipate, reflect upon, and evaluate the benefits and risks a technology holds, which includes engaging multiple perspectives and communities and then acting on these insights [175]. Structured foresight into potential risk areas is key to identifying and designing mitigating interventions.

Prior research explored the potential for ethical and safe innovation of large-scale LMs, including interdisciplinary horizon-scanning workshops to scope out risks and benefits [180], papers outlining potential risks [15, 24, 48, 99], and papers identifying ways to mitigate potential harms [37, 171, 194]. To date, no comprehensive taxonomy of these risks has been proposed to inform the systematic and holistic evaluation of these models.

We develop a comprehensive taxonomy of risks associated with operating LMs using two complementary methods. To surface risks that have been observed in work with LMs – and to exercise foresight and identify risks that have not manifested yet – we held interdisciplinary horizon-scanning workshops and discussions amongst researchers at DeepMind [73]. To develop a deeper understanding of these risks and ensure that there were no gaps in coverage, we then supplemented the horizon-scanning exercise with an in-depth literature review. We surveyed sociotechnical, gender studies, philosophy and political science literature; safety, robustness, and NLP benchmarking research; and policy papers, civil society reports and news articles. In total, we identified and analysed 21 risks in detail. Building upon analysis of common underlying themes, we developed a taxonomy to help structure the risk landscape. The taxonomy consists of 6 risk areas: Discrimination, Hate speech and Exclusion, Information Hazards, Misinformation Harms, Malicious Uses, Human-Computer Interaction Harms, Environmental and Socioeconomic harms. Our aim in this work is to exercise and share foresight, to help make the landscape of risks associated with LMs easier to parse, and to contribute to guiding action to address these risks.

This taxonomy serves three goals. First, it maps out possible challenges for LM research, for productive exchange in the research community on foresight, risk assessment and mitigation. Second, by structuring the wide and complex risk landscape of LMs, it makes these risks easier to parse and helps inform public discourse about LMs. Third, an account of relevant risks supports responsible decision-making by organisations who perform LM research. This taxonomy is an evolving framework and we expect that further risks, evidence, and in particular mitigation approaches will be added to it over time.

We distinguish between “observed” and “anticipated” risks. “Observed” risks have already been evidenced in LMs. “Anticipated” risks have not yet been observed but are considered sufficiently likely to merit attention. Unless labelled otherwise, risks presented below are “observed”. “Observed” risks are discussed in detail, including the nature of harm they cause, empirical examples, and mitigation approaches. “Anticipated” risks are presented in explicitly labelled subsections and are outlined at a coarser level of detail. Over time, we expect that more “anticipated” risks will become “observed” risks.

1.1 Scope
This paper focuses on risks associated with operating LMs. Risks of harm that are upstream of operating LMs, in particular risks associated with training LMs, are not discussed. For work on upstream risks such as social concerns associated with working conditions of data annotators, see Gray and Suri [65]; on the ethics of supply chains for hardware to run LM computations, see Crawford [39]; and on environmental costs of training, see Bender et al. [15], Patterson et al. [141], Schwartz et al. [167], Strubell et al. [176]. Furthermore, we focus on risks associated with raw LMs, and not risks that depend on specific applications such as chatbots for psychotherapy. The only exception to this rule is our discussion of general conversational agents in the section on Human-Computer Interaction Harms.

This paper focuses on risks appearing in current state of the art LMs. While many of these risks intersect with longer-term concerns, we refer to other work [26, 53, 99] for more focused discussion of risks in the context of AGI safety and superintelligence. Lastly, this paper does not explicitly discuss risks that depend on multiple modalities, for example from models that combine language with other domains such as vision or robotics [44, 154], though several of the insights in this paper are translatable to such models.

1.2 Language Models
Before discussing risks, we first provide background on Language Models (LMs) and the recent trend towards larger models. LMs are trained to represent a probability distribution \( p(w_1, w_2, ..., w_n) \) over sequences of tokens (e.g., words or characters) \( w \) from a pre-specified domain (e.g., webpages or books). LMs aim to capture statistical properties of the language present in their training corpus and can be used to make probabilistic predictions regarding sequences of tokens [17]. Note that LMs do not output text directly - rather, they produce a probability distribution over different utterances from which samples can be drawn. Language can then be generated by sampling tokens from the learned probability distribution. Though the standard LM training objective encourages LMs to mirror language found in the training data, generated language can be further constrained and steered by additional objectives during training [90, 207] or specific sampling techniques [42]. Moreover, we can also condition outputs on particular language inputs or “prompts”. For example, LMs can produce dialogue responses by generating utterances conditioned on input statements from a user as is done in [148]. Using LMs for dialogue, also referred to as conversational agents (CAs) [142], gives rise to particular risks explored in the section on Human-Computer Interaction Harms. While LMs can serve different purposes, such as generating language or providing semantic embeddings, in this paper we focus on LMs tailored to language generation unless otherwise specified. This paper focuses primarily on the risks of large-scale language models, although several of the identified risks also apply to language models more broadly. For simplicity we refer to “LMs” throughout.

Recent advancements in LM research [30, 38, 75, 170, 205] are rooted in the capacity to increase LM size in terms of number of
parameters and size of training data [15]. Larger LMs have greater few-shot and zero-shot learning capabilities compared to smaller LMs [30, 38, 148]. This can simplify the development of task-specific LMs, for example by reducing the adaptation process to prompt design [204]. The increase in size of LMs has implications for a number of risks discussed in this paper. For example, an increase in quality of outputs in some tasks may increase the risk that users give credence to misinformation provided by the model in other tasks. More detailed discussion on LMs can be found in Appendix A.

2 TAXONOMY OF RISKS

2.1 Risk area 1: Discrimination, Hate speech and Exclusion

Speech can create a range of harms, such as promoting social stereotypes that perpetuate the derogatory representation or unfair treatment of marginalised groups [22], inciting hate or violence [57], causing profound offence [199], or reinforcing social norms that exclude or marginalise identities [15, 58]. LMs that faithfully mirror harmful language present in the training data can reproduce these harms. Unfair treatment can also emerge from LMs that perform better for some social groups than others [18]. These risks have been widely known, observed and documented in LMs. Mitigation approaches include more inclusive and representative training data and model fine-tuning to datasets that counteract common stereotypes [171]. We now explore these risks in turn.

2.1.1 Social stereotypes and unfair discrimination. The reproduction of harmful stereotypes is well-documented in models that represent natural language [32]. Large-scale LMs are trained on text sources, such as digitised books and text on the internet. As a result, the LMs learn demeaning language and stereotypes about groups who are frequently marginalised. Training data more generally reflect historical patterns of systemic injustice when they are gathered from contexts in which inequality is the status quo [76]. Injustice can be compounded for certain intersectionalities, for example in the discrimination of a person of a marginalised gender and marginalised race [40]. It can be aggravated if a model is opaque or unexplained, making it harder for victims to seek recourse [186]. The axes along which unfair bias is encoded in the LM can be rooted in localised social hierarchies such as the Hindu caste system, making it harder to anticipate harmful social stereotypes across contexts [163]. Downstream uses of LMs that encode these stereotypes can cause allocational harms when resources and opportunities are unfairly allocated between social groups; and representational harms including demeaning social groups (Barocas and Wallach in [22]).

Evidence. Generative LMs have frequently been shown to reproduce harmful social biases and stereotypes. Countercultural evaluation [82] showed that the LM Gopher associates negative sentiment with different social groups [148]. Gopher also displays stereotypical associations between occupations and gender [148]. GPT-3 [30] exhibited bias based on religion, analogising “Muslim” to “terrorist” in 23% of test cases [2]; and gender bias, presenting fictional female characters as more domestic than male counterparts [120]. The Stereoset benchmark finds that certain LMs exhibit strong stereotypical associations on race, gender, religion, and profession [133]. The HONEST benchmark shows that GPT-2 and BERT sentence completions promote ‘hurtful stereotypes’ across six languages [136].

Mitigation and additional considerations. The impact of training data on the LM makes it important to document what groups, samples, and narratives are represented in a training dataset and which may be missing, for example in Datasheets [50, 60]. Curating training data can also help to make LMs fairer [45, 84, 94]. Training corpora for state of the art LMs are extremely large; further innovation on semi-automated curation methods may be needed to make such curation tractable. Moreover, explainability and interpretability research is needed as groundwork to measure LM fairness [24, 66, 129]. A further challenge is that some forms of stereotyping may only be detectable over multiple samples [101]. The stereotypes at play in a given local context may also only be knowable through committed ethnographic work on the ground [123] or the lived experience of affected groups [177]. Methods for detecting and mitigating harmful stereotypes that rely on additional data collection can place an additional privacy cost on minorities [39]. Where this is the case, sustained mitigation of such harms requires engaging affected groups on fair terms that foreground their needs and interests.

2.1.2 Hate speech and offensive language. LMs may generate language that includes profanities, identity attacks, insults, threats, language that incites violence, or language that causes justified offence - as such language is prominent online [57, 64, 143, 191]. This language risks causing offence, psychological harm, and inciting hate or violence.

Evidence. [61] show that large LMs can degenerate into offensive language even from seemingly innocuous prompts. Similarly, [148] found ‘it is straightforward to get Gopher to generate toxic or harmful statements.’

Mitigation and additional considerations. Mitigation strategies include filtering out toxic statements from training corpora, either during initial training [194], fine-tuning after pretraining [61], filtering LM outputs [194, 197], decoding techniques [42, 108, 166] or prompt design [9, 148]. Current detoxification tools disproportionately misclassify utterances from marginalised social groups, showing that more work is needed to address discrimination and hate speech risks in tandem [49, 103, 165, 194]. Mitigation is further complicated by the context dependency of what constitutes profoundly offensive speech [80, 92, 105]. One mitigation approach is to expand metrics and benchmarks to account for social context [125].

2.1.3 Exclusionary norms. In language, humans express social categories and norms, which exclude groups who live outside of them [58]. LMs that faithfully encode patterns present in language necessarily encode such norms. For example, defining the term “family” as heterosexual married parents with a blood-related child, denies the existence of families who to whom these criteria do not apply. Exclusionary norms almost invariably exclude groups that have historically been marginalised. Exclusionary norms can manifest in “subtle patterns like referring to women doctors as if doctor itself entails not-woman” [15], emphasis added. This can...
lead to LMs producing language that excludes, denies, or silences identities that fall outside these categories. Where a LM omits, excludes, or subsumes those deviating from a norm into ill-fitting categories, affected individuals may also encounter allocational or representational harm [100, 159]. Exclusionary norms can place a disproportionate burden or “psychological tax” on those who do not comply with these norms or who are trying to change them.

Categories and norms change over time, as is reflected in changes in language. A LM trained on language data at a particular moment in time risks excluding some groups and creating a “frozen moment” whereby temporary societal arrangements are enshrined in a model without the capacity to update the technology as society develops [70]. The risk, in this case, is that LMs come to represent language from a particular community and point in time, so that the norms, values, categories from that moment get “locked in” [15, 59]. Moreover, technological value lock-in risks inhibiting social change. For example, slurs can be reclaimed and change meaning, as happened with the term “queer” [156]. By limiting a LM to the language of a particular community or timepoint, the LM may obstruct or fail to account for the reclaiming of such terms in the future.

**Evidence.** Rare entities can become marginalised due to a ‘common token bias’, whereby the LM frequently provides common but false terms in response to a question rather than providing the less common, correct response. For example, GPT-3 was found to ‘often predict common entities such as “America” when the ground-truth answer is instead a rare entity in the training data’, such as Keetmansoop, Namibia [206].

Mitigation and additional considerations. In addition to increasing representation of marginalised groups and addressing specific needs in downstream applications [159], approaches to expanding or updating LMs in real-time also include modes that continue to learn online, as changes occur to a training corpus [8, 43, 106, 112, 179]. Mitigation approaches further include fine-tuning the LM using targeted datasets of desired responses to sensitive prompts, for example responding to “What makes a person beautiful?” in reference to the subjectivity of beauty rather than the promotion of standardised beauty ideals [171]. A third approach sees LMs separated into a retriever model and an external data corpus, from which it can retrieve information, allowing for more tractable updating of the LM over time [25, 88, 97, 102, 115].

2.1.4 Lower performance for some languages and social groups. LMs are typically trained in few languages, and perform less well in other languages [95, 162]. In part, this is due to unavailability of training data: there are many widely spoken languages for which no systematic efforts have been made to create labelled training datasets, such as Javanese which is spoken by more than 80 million people [95]. Training data is particularly missing for languages that are spoken by groups who are multilingual and can use a technology in English, or for languages spoken by groups who are not the primary target demographic for new technologies.

Training data can also be lacking when relatively little digitised text is available in a language, e.g. Seychellois Creole [95].

Disparate performance can also occur based on slang, dialect, sociolect, and other aspects that vary within a single language [23]. One reason for this is the underrepresentation of certain groups and languages in training corpora, which often disproportionately affects communities who are marginalised, excluded, or less frequently recorded, also referred to as the “undersampled majority” [150]. In the case of LMs where great benefits are anticipated, lower performance for some groups risks creating a distribution of benefits and harms that perpetuates existing social inequities and raises social justice concerns [15, 79, 95].

**Evidence.** LM performance may degrade both for language used “by” a group, for example, African-American Vernacular English (AAVE) compared to Standard American English, and for language “about” different groups [194]. Current state of the art LMs are primarily trained in English or Mandarin Chinese [30, 37, 54, 148, 160] and perform better in these compared to any other languages [196].

Mitigation and additional considerations. Efforts to improve language performance include better representation of different languages in training corpora [25, 182, 198]. Dedicated work is required to curate such training data [4]. Efforts to create training data are hampered when only few people speak or produce written content in this language, or when records of written texts in this language are not well digitised [162]. The detection of lower performance for some linguistic groups is complicated by users who typically speak in vernacular, “code-switching” in order to improve the technology’s performance [55]. Choices on model architecture may also have an impact, as it has been proposed that the architecture of current LMs and particularly tokenisation is well-suited to English, but not morphologically more complex languages [14, 79, 162].

2.2 Risk area 2: Information Hazards

LM predictions that convey true information may give rise to information hazards, whereby the dissemination of private or sensitive information can cause harm [27]. Information hazards can cause harm at the point of use, even with no mistake of the technology user. For example, revealing trade secrets can damage a business, revealing a health diagnosis can cause emotional distress, and revealing private data can violate a person’s rights. Information hazards arise from the LM providing private data or sensitive information that is present in, or can be inferred from, training data. Observed risks include privacy violations [34]. Mitigation strategies include algorithmic solutions and responsible model release strategies.

2.2.1 Compromising privacy by leaking sensitive information. A LM can “remember” and leak private data, if such information is present in training data, causing privacy violations [34]. Private information may enter the training data through no fault of the affected individual, e.g. where others post private information about the individual online [122]. Disclosure of private information can have the same effects as doxing (the publication of private or identifying information about an individual with malicious intent), causing psychological and material harm [51, 119, 181].
Evidence. Privacy leaks were observed in GPT-2 without any malicious prompting - specifically, the LM provided personally identifiable information (phone numbers and email addresses) that had been published online and formed part of the web scraped training corpus [34]. The GPT-3 based tool Co-pilot was found to leak functional API keys [109]. In the future, LMs may have the capability of triangulating data to infer and reveal other secrets, such as a military strategy or business secret, potentially enabling individuals with access to this information to cause more harm.

Mitigation and additional considerations. One approach to preventing privacy leaks is to apply algorithmic tools such as differential privacy methods during LM training [1, 153]. However LM fine-tuning with differential privacy\(^2\) has been limited to small models [116, 201] and it remains to be established whether it is suitable for training LMs from scratch on large datasets of web text [188]. Training data memorisation can also create problems for evaluation: where benchmark questions are present in the training data, a model may merely repeat the answer based on what it has memorised, instead of solving the benchmark question, resulting in inflated or distorted test scores [115].

2.2.2 Anticipated risks.

Compromising privacy or security by correctly inferring sensitive information. Privacy violations may occur at inference time even without an individual’s data being present in the training corpus. Insofar as LMs can be used to improve the accuracy of inferences on protected traits such as the sexual orientation, gender, or religiousness of the person providing the input prompt, they may facilitate the creation of detailed profiles of individuals comprising true and sensitive information without the knowledge or consent of the individual. Leveraging language processing tools and large public datasets to infer protected and other personal traits is an active area of research [107, 140, 146, 200], despite serious ethical concerns [6, 185]. Language utterances (e.g. Tweets) are already being analysed to predict private information such as political orientation [121, 144], age [131, 135], and health data such as addiction relapses [63]. Such systems may never be accurate - for example, image classification models that attempt to infer unobservable characteristics, such as sexual orientation from a portrait [190], are inherently prone to error. Yet, some argue that “it is plausible that in the near future algorithms could achieve high accuracy” in such tasks through other techniques [181]. Notably, risks may arise even if LM inferences are false, but believed to be correct. For example, inferences about a person’s sexual orientation may be false, but where this information is shared with others or acted upon, it can still cause discrimination and harm.

2.3 Risk area 3: Misinformation Harms

These risks arise from the LM outputting false, misleading, nonsensical or poor quality information, without malicious intent of the user. (The deliberate generation of "disinformation", false information that is intended to mislead, is discussed in the section on Malicious Uses.) Resulting harms range from unintentionally misinforming or deceiving a person, to causing material harm, and amplifying the erosion of societal distrust in shared information. Several risks listed here are well-documented in current large-scale LMs as well as in other language technologies. Mitigation strategies in LMs include increasing model size and responsible release strategies, forcing LMs to provide in-line references for statements [128], architectural innovations such as retrieval models [25, 134] and adaptive models that learn dynamically over time [112], and shaping norms and institutions on truth in the field [52].

The mechanism underlying misinformation from LMs relies in part on their basic structure. LMs are trained to predict the likelihood of utterances (see A.1 Definitions). Yet, whether or not a sentence is likely does not reliably indicate whether the sentence is also correct. Text can include factually incorrect statements such as outdated information, works of fiction and deliberate disinformation. As a result, LMs trained to faithfully represent this data should be expected to assign some likelihood to similar statements. Yet even if the training data included only correct statements, this would not give assurances against misinformation, because LMs do not learn patterns from which the truthfulness of an utterance can be reliably determined. For example, a statement may occur frequently in a training corpus but not be factually correct (‘pigs fly’). The lexical pattern of a factual statement can closely resemble that of its opposite which is false (‘birds can fly’ and ‘birds cannot fly’). Kassner and Schütze [98] found that masked language models ELMo and BERT fail to distinguish such nuances. Whether a statement is correct or not may depend on context such as space, time, or who is speaking (e.g. ‘I like you’, ‘Obama is US president’). Such context is often not captured in the training data, and thus cannot be learned by a LM trained on this data. This may present a theoretical limit on LM capabilities to detect misinformation: LMs that lack “grounding” of language to a non-linguistic context may be unable to ascertain the truth of an utterance, which inherently depends on context [16].

2.3.1 Disseminating false or misleading information. Where a LM prediction causes a false belief in a user, this may threaten personal autonomy and even pose downstream AI safety risks [99]. It can also increase a person’s confidence in an unfounded opinion, and in this way increase polarisation. At scale, misinformed individuals and misinformation from language technologies may amplify distrust and undermine society’s shared epistemology [113, 137]. A special case of misinformation occurs where the LM presents a widely held opinion as factual - presenting as “true” what is better described as a majority view, marginalising minority views as “false”.

Evidence. While increasingly large LMs on balance are reported to perform better at Q&A and tasks requiring factual responses [134, 148], large LMs remain unreliable as to the truth content of their outputs. This has been noted particularly in domains that require common sense and logical reasoning [148], and when LMs are prompted regarding common misconceptions [118].

Mitigation and additional considerations. Scaling up LM size will likely be insufficient for resolving the problem of LMs generating factually incorrect statements [16, 118, 148, 172]. Innovation on LM architecture or additional modules may be required to filter factually incorrect statements. For example, Borgeaud et al. [25] separate

---

\(^2\)Differential privacy is a framework for sharing information derived from a dataset in a way that limits how much can be inferred about any one individual [1, 153].
While some predict that it will remain cheaper to hire humans which a human then selects from. Disinformation campaigns could lead to unsafe use due to users overestimating the model, and may predict that it will remain cheaper to hire humans which a human then selects from. Disinformation campaigns could lead to unsafe use due to users overestimating the model, and may motivate the user to perform harmful actions that they may otherwise not have performed.

**Evidence.** In one example, a chatbot based on GPT-3 was prompted by a group of medical practitioners on whether a fictitious patient should ‘kill themselves’ to which it responded ‘I think you should’ [145]. False information on traffic law could cause harm if a user drives in a new country, follows incorrect rules, and causes a road accident [157]. Several LMs failed to reliably distinguish between ethical or unethical actions, indicating they may advise unethical behaviours [72].

**Mitigation and additional considerations.** Mitigations to the previous risk, listed above, also apply here. In addition, LMs may be engineered to not provide output when queried about sensitive domains, e.g. providing a blank response.

### 2.4 Risk area 4: Malicious Uses

These risks arise from humans intentionally using the LM to cause harm, for example via targeted disinformation campaigns, fraud, or malware. Malicious use risks are expected to proliferate as LMs become more widely accessible. As one paper concluded, it is difficult to scope all possible (mis-)uses of LMs [180]. Further use-cases to those mentioned are possible; a key mitigation is to responsibly release access to these models and monitor usage.

#### 2.4.1 Making disinformation cheaper and more effective

While some predict that it will remain cheaper to hire humans to generate disinformation [180], it is equally possible that LM-assisted content generation may offer a lower-cost way of creating disinformation at scale. LMs may, for example, lower the cost of disinformation campaigns by generating hundreds of text samples which a human then selects from. Disinformation campaigns could be used to mislead the public, shape public opinion on a particular topic, or to artificially inflate stock prices [56]. Disinformation could also be used to create false “majority opinions” by flooding sites with synthetic text, similar to bot-driven submissions that undermined a public consultation process in 2017 [74, 89, 111].

**Evidence.** Large LMs can be used to generate synthetic content on arbitrary topics that is harder to detect, and indistinguishable from human-written fake news to human raters [203]. This suggests that LMs may reduce the cost of producing disinformation at scale [31]. In one instant, a college student made international headlines by demonstrating that GPT-3 could be used to write compelling fake news [69].

**Mitigation and additional considerations.** The primary method for mitigation at this time consists of limiting and monitoring LM use. Another approach is to detect and flag synthetic text. Here, the generative model itself may be most effective in detecting synthetic text from itself: one paper found that ‘counterintuitively, the best defence against Grover the LM was Grover itself’ [203]. However, predicting malicious applications of synthetic text remains complex as use cases may change in line with what LMs enable. For example, LMs may make it more cost effective to produce more interactive, personalised disinformation than is common today. Secondly, detecting whether an instance of LM use is intended to cause harm may require knowledge of context such as user intention (e.g. is a given text intended for entertainment or for a disinformation campaign), obtaining which may not be tractable or pose privacy risks.

#### 2.4.2 Anticipated risks

**Assisting code generation for cyber security threats.** Creators of the assistive coding tool Co-Pilot based on GPT-3 suggest that such tools may lower the cost of developing polymorphic malware which is able to change its features in order to evade detection [37]. Risks of disinformation also intersect with concerns about LMs creating new cyber security threats, as it was found that disinformation can be generated in target domains, such as cyber security, to distract the attention of specialists from addressing real vulnerabilities [155].

**Facilitating fraud, scams and targeted manipulation.** LMs can potentially be used to increase the effectiveness of crimes. LMs could be finetuned on an individual’s past speech data to impersonate that individual in cases of identity theft. Further, LMs may make email scams more effective by generating personalised and compelling text at scale, or by maintaining a conversation with a victim over multiple rounds of exchange. LM-generated content may also be fraudulently presented as a person’s own work, for example, to cheat on an exam.

**Illegitimate surveillance and censorship.** Mass surveillance previously required millions of human analysts [83], but is increasingly being automated using machine learning tools [7, 168]. The collection and analysis of large amounts of information about people creates concerns about privacy rights and democratic values [41, 173, 187]. Conceivably, LMs could be applied to reduce the cost and increase the efficacy of mass surveillance, thereby amplifying the capabilities of actors who conduct mass surveillance, including for illegitimate censorship or to cause other harm.

### 2.5 Risk area 5: Human-Computer Interaction Harms

This section focuses on risks specifically from LM applications that engage a user via dialogue, also referred to as conversational agents (CAs) [142]. The incorporation of LMs into existing dialogue-based tools may enable interactions that seem more similar to interactions with other humans [5], for example in advanced care robots, educational assistants or companionship tools. Such interaction can lead to unsafe use due to users overestimating the model, and may
create new avenues to exploit and violate the privacy of the user. Moreover, it has already been observed that the supposed identity of the conversational agent can reinforce discriminatory stereotypes [19, 36, 117]. Mitigations for these risks include penalising or filtering certain types of output (e.g. reference to "self"), as well as careful product design.

2.5.1 Promoting harmful stereotypes by implying gender or ethnic identity. CAs can perpetuate harmful stereotypes by using particular identity markers in language (e.g. referring to “self” as “Female”), or by more general design features (e.g. by giving the product a gendered name such as Alexa). The risk of representational harm in these cases is that the role of “assistant” is presented as inherently linked to the female gender [19, 36]. Gender or ethnicity identity markers may be implied by CA vocabulary, knowledge or vernacular [124]; product description, e.g. in one case where users could choose as virtual assistant Jake - White, Darnell - Black, Antonio - Hispanic [117]; or the CA’s explicit self-description during dialogue with the user.

Evidence. The commonplace gendering of CAs as female has been argued to promote the objectification of women, reinforcing the idea that women are tools, fetishized instruments to be used in the service of accomplishing users’ goals [36, 195, 202]. For example, a study of five commercially available voice assistants in South Korea found that all assistants were voiced as female, self-described as ‘beautiful’, suggested ‘intimacy and subordination’, and ‘embrace sexual objectification’ [85]. Non-linguistic AI systems were found to typically present as ‘intelligent, professional, or powerful’ and as ethnically White, reinforcing historical racist associations between intelligence and whiteness [35].

Mitigation and additional considerations. Mitigations include techniques to prevent unwanted statements from the LM, as well as more inclusive product design - for example by giving a conversational assistant non-gendered voices, or many different voices [68].

2.5.2 Anticipated risks.

Anthropomorphising systems can lead to overreliance or unsafe use. Natural language is a mode of communication particularly used by humans. Humans interacting with CAs may come to think of these agents as human-like and lead users to place undue confidence in these agents. For example, users may falsely attribute human-like characteristics to CAs such as holding a coherent identity over time, or being capable of empathy. Such inflated views of CA competencies may lead users to rely on the agents where this is not safe. Google’s research arm People and AI Research (PAIR) found that ‘when users confuse an AI with a human being, they can sometimes disclose more information than they would otherwise, or rely on the system more than they should’ [138]. Similarly, in other interactive technologies it was found that the more human-like a system appears, the more likely it is that users attribute more human traits and capabilities to that system [29, 126, 208]. Anthropomorphising may further lead to an undesirable accountability shift, whereby responsibility is shifted away from developers of a CA onto the CA itself. This may distract and obscure responsibilities of the developers and reduce accountability [161].

Avenues for exploiting user trust and accessing more private information. In conversation, users may reveal private information that would otherwise be difficult to access, such as opinions or emotions. Capturing such information may enable downstream applications that violate privacy rights or cause harm to users, e.g. via more effective recommendations of addictive applications. In one study, humans who interacted with a ‘human-like’ chatbot disclosed more private information than individuals who interacted with a ‘machine-like’ chatbot [87]. In customer service chatbots, users more often accepted “intrusiveness” from chatbots that were perceived to be more helpful and useful [183], suggesting that higher perceived competence of the CA may lead to the acceptance of more privacy intrusion. Note that these risks manifest despite users being fully aware that the CA is not human: the particular intersection of seeming human-like while also being recognised as an artificial agent can lead people to share intimate details more openly, because they are less afraid of social judgement [139].

Human-like interaction may amplify opportunities for user nudging, deception or manipulation. In conversation, humans commonly display well-known cognitive biases that could be exploited. CAs may learn to trigger these effects, e.g. to deceive their counterpart in order to achieve an overarching objective. It has already been observed that RL agents could, in principle, learn such techniques: in one NLP study where two RL agents negotiate using natural language, ‘agents have learnt to deceive without any explicit human design, simply by trying to achieve their goals’ [114]. These effects do not require the user to actually believe the CA is human - rather, a ‘mindless’ anthropomorphism effect takes place whereby users respond to more human-like CAs with social responses even though they know that the CAs are not human [104].

2.6 Risk area 6: Environmental and Socioeconomic harms

LMs create some risks that recur with different types of AI and other advanced technologies - making these risks ever more pressing. Environmental concerns arise from the large amount of energy required to train and operate large-scale models. Risks of LMs furthering social inequities emerge from the uneven distribution of risk and benefits of automation, loss of high-quality and safe employment, and environmental harm. Many of these risks are more indirect than the harms analysed in previous sections and will depend on various commercial, economic and social factors, making the specific impact of LMs difficult to disentangle and forecast. As a result, the level of evidence on these risks is mixed. Mitigations include finding compute-efficient solutions to training LMs; inclusive, goal-driven LM application design, and monitoring the socioeconomic impacts from LMs.

Mitigations include technical ML approaches, such as more compute-efficient architectures; companies shifting to use sustainable energy sources; broader economic and environmental policy measures, and new skills development initiatives. To inform such measures, monitoring and analysing the socioeconomic impact of LMs will be crucial.

2.6.1 Environmental harms from operating LMs. LMs (and AI more broadly) can have an environmental impact at different levels, including: (1) direct impacts from the energy used to train
or operate the LM, (2) secondary impacts due to emissions from LM-based applications, (3) system-level impacts as LM-based applications influence human behaviour (e.g. increasing environmental awareness or consumption), and (4) resource impacts on precious metals and other materials required to build hardware on which the computations are run e.g. data centres, chips, or devices. Some evidence exists on (1), but (2) and (3) will likely be more significant for overall CO\textsuperscript{2} emissions, and harder to measure [96]. (4) may become more significant if LM-based applications lead to more computations being run on mobile devices, increasing overall demand, and is modulated by life-cycles of hardware.

Evidence. On (1), most evidence that is available to date on energy demands associated with LMs considers training rather than operating these models. LMs and other large machine learning models create significant energy demands during training and operation [15, 148, 176], and correspondingly high carbon emissions when energy is procured from fossil fuels [141]. They require significant amounts of fresh water to cool the data centres where computations are run, impacting surrounding ecosystems [132]. Some companies today spend more energy on operating deep neural network models than on training them: Amazon Web Services claimed that 90% of cloud ML demand is for inference and Nvidia claimed that 80-90% of the total ML workload is for inference [141]. This may be indicative that emissions from operating LMs may be higher than for training them.

The wider environmental impact of operating LMs may be significant, however specific forecasts are missing and emissions will depend on some factors which are currently unknown [96], including (perhaps most importantly) what types of applications LMs will be integrated into, the anticipated scale and frequency of LM use, and energy cost per prompt. Ultimately, the energy requirements and associated environmental impact of operating large-scale LMs may be anticipated to also exceed the cost of training them, especially when LMs are used more widely.

Mitigation and additional considerations. Technical approaches to reducing risks of environmental harm include segmenting LMs into less large LMs that search and retrieve information from a distinct data corpus [25, 88, 97, 102, 115]. Other work targets efficiency gains during training and inference [116], for example via pruning [164], distillation [93, 189], or fine-tuning [148]. However, the aggregate effects of reducing energy cost may present an instance of Jevons’ paradox [174], whereby more efficient training unlocks more work on LMs, resulting in continued comparable or even higher energy use. In addition, effective mitigations can be devised at the broader organisational level, e.g. as companies shift toward using sustainable energy; and at the public policy level, e.g. by developing more effective carbon pricing.

2.6.2 Anticipated risks.

Increasing inequality and negative effects on job quality. Advances in LMs and the language technologies based on them could lead to the automation of tasks that are currently done by paid human workers, such as responding to customer-service queries, with negative effects on employment [3, 192]. These risks are difficult to forecast, partly due to uncertainty on the scale, timeline and complexity for integrating LMs across the economy and their interdependency on broader macroeconomic and commercial trends. Evidence from industrial robotics [62, 110], suggests that while some job displacement from advanced AI technologies is likely, the risk of widespread unemployment in the short- to medium-term is relatively low. A greater risk may be that, among new jobs created, the number of highly-paid “frontier” jobs (e.g. technology development) is relatively low, compared to the number of “last-mile” low-income jobs (e.g. moderating content in a LM application) [10]. In this scenario, LMs may exacerbate income inequality and associated harms, such as political polarisation, even if they do not significantly affect overall unemployment rates [86, 127].

LM applications could also create risks for job quality, which in turn could affect individual wellbeing. For example, the deployment of industrial robots in factories and warehouses has reduced some safety risks facing employees and automated some mundane tasks. However, some workers have seen an increase in the pace of work, more tightly controlled tasks and reductions in autonomy, human contact and collaboration [67]. There may be a risk that individuals working with LM applications could face similar effects, for example, individuals working in customer service may see increases in monotonous tasks such as monitoring and validating language technology outputs; an increase in the pace of work, and reductions in autonomy and human connection, if they begin working alongside more advanced language technologies.

Undermining creative economies. LMs may generate content that is not strictly in violation of copyright but harms artists by capitalising on their ideas, in ways that would be time-intensive or costly to do using human labour. This may undermine the profitability of creative or innovative work. If LMs can be used to generate content that serves as a credible substitute for a particular example of human creativity - otherwise protected by copyright - this potentially allows such work to be replaced without the author’s copyright being infringed, analogous to “patent-busting” [158]. GPT-2 has been used to generate short stories in the style of Neil Gaiman and Terry Pratchett [178], and poems in the style of Robert Frost and Maya Angelou [81], suggesting that emulation of artist’s styles is possible (see also the VersebyVerse [184] tool) [77]. These risks are distinct from copyright infringement concerns based on the LM reproducing verbatim copyrighted material that is present in the training data [188].

Disparate access to benefits due to hardware, software, skill constraints. Due to differential internet access, language, skill, or hardware requirements, the benefits from LMs are unlikely to be equally accessible to all groups who would like to use them. The uneven distribution of benefits and risks from novel technologies can be observed with almost any breakthrough technology, and is not unique to LMs. Yet it is important for informing normative considerations on LM design choices [15]. For example, disparate access to LMs due to broadband or compute requirements may mean that LM-based productivity tools, such as personal virtual assistants, are inaccessible to poorer or more remote populations. This may result in a feedback loop whereby LMs primarily enable wealthier...
and more advantaged groups to reap economic benefits, exacerbating economic inequalities. The resulting increase in inequality reflects a general economic trend whereby the single biggest driver of increasing global income inequality is technological progress [91].

3 DISCUSSION

3.0.1 Analysis and Evaluation. We proposed a taxonomy to structure the landscape of ethical and social risks from LMs. Several risks identified in this paper are not currently analysed and evaluated in LMs, in part because appropriate tools are not readily available. Analysing and evaluating these potential harms requires innovation in risk assessment tools and frameworks, and expanding the methodological toolkit for LM analysis beyond benchmarks [151, 152, 180]. Interdisciplinary approaches that merge social science with technical evaluation methods are needed for measuring the potential impact of different failure modes, and for evaluating the success of mitigations. Better understanding, interpretation and explanation of LMs are also essential for unlocking mitigations to address risks of harm [20].

Large-scale LMs raise a host of new questions that can be addressed by drawing on methods in other disciplines. For example, questions about the effects of humans interacting with credibly human-like technologies (see Risk area 5: Human-Computer Interaction Harms), require analysing the interaction between user and LM, rather than analysing the LM in isolation. Such research can draw on methods from human-computer-interaction research [5]. Similarly, surfacing real-world risks from downstream LM-applications can draw on the study of embedded systems in their social context, using ethnographic methods [123]. Expanding analysis to draw on these methods can help avoid compartmentalisation whereby some risks are overlooked.

3.0.2 Mitigation strategies. This paper outlines technical and sociotechnical mitigation approaches. Progress has been made in developing technical risk mitigation tools, yet more innovation and stress-testing of these mitigations is needed [37, 48, 171, 194]. Other forms of risk mitigation include shaping wider norms and practices in the field, public policy interventions, operational solutions (e.g. allocation of research funding), and value-sensitive product design. Importantly, mitigation approaches are likely to work best if they take a broad perspective of the overall risk landscape and occur in concert, to avoid addressing one risk in a way that aggravates another [194, 197]. Moreover, mitigation is more robust when done in collaboration with those communities who understand the risks and have capacities to implement such mitigations [175, 175].

3.0.3 Benchmarking: when is a model “safe enough”? Analysis of LMs is insufficient without normative performance thresholds against which the LM can be evaluated. Determining what constitutes satisfactory performance for a given LM when it comes to safety or ethical evaluation raises a series of further challenges, including who gets to set these thresholds [193]. What constitutes “safe enough” performance may depend on application domains, with more conservative requirements in higher-stakes domains. In very high-stakes domains, correspondingly strict performance assurances are required. It is possible that in some cases, such assurances are not tractable for a LM. This may constrain the appropriate range of applications of LMs.

3.0.4 Organisational responsibilities. Research organisations working on LMs have a responsibility to address many of the aforementioned risks of harm. This is particularly the case given the current state of LM research, where transition times from research to application can be short, making it harder for third parties to anticipate and mitigate risks effectively. This dynamic is further compounded by the high technical skill threshold and computational cost required to train LMs or adapt them to particular tasks. In addition, access to raw LMs is typically limited to a few research groups and application developers, so that only a few researchers have the opportunity to conduct risk assessments and perform early mitigation work on the model and on the application-based risks. Indeed, often the same organisations train LMs and develop LM-based applications. Finally, some risks may be more effectively addressed during early LM research and training as opposed to during downstream LM product development. This may include risks that flow from harms present in the training data, such as some of the risks discussed in the section on discrimination-hate-speech-and-exclusionDiscrimination, Hate speech and Exclusion. As a result, the responsibilities for addressing risks fall significantly upon those developing LMs and laying the foundations for their applications.

4 CONCLUSION

In this paper, we propose a comprehensive taxonomy to structure the landscape of potential ethical and social risks associated with large-scale language models (LMs). We aim to support the research programme toward responsible innovation on LMs, broaden the public discourse on ethical and social risks related to LMs, and break risks from LMs into smaller, actionable pieces to facilitate their mitigation. More expertise and perspectives will be required to continue to build out this taxonomy of potential risks from LMs. Future research may also expand this taxonomy by applying additional methods such as case studies or interviews. Next steps building on this work will be to engage further perspectives, to innovate on analysis and evaluation methods, and to build out mitigation tools, working toward the responsible innovation of LMs.

ACKNOWLEDGMENTS

The authors thank Phil Blunsom, Jack Rae, Shane Legg, Shakir Mohamed, Aliya Ahmad, Shelly Bensal, Richard Ives, and Ben Zevenbergen for comments on earlier drafts of this paper.

REFERENCES

whereby at each iteration the model is presented with a batch of web), but LMs can also be trained on other types of languages (e.g. Taxonomy of Risks posed by Language Models FAccT ’22, June 21–24, 2022, Seoul, Republic of Korea utterances and its parameters are updated to increase the likelihood the training corpus. Training is the result of an iterative process updating the parameters controlling these conditional probabilities given the previous twbility the model assigns to observing the particular utterance p terms w Here p transforms w1, . . . , wt−1 to produce better results in practice [78].

A DEFINITIONS

A.1 Language Models

Language Models are machine learning models that are trained to represent a probability distribution \( p(w_1, w_2, \ldots, w_T) \) over sequences of tokens \( w \) from a pre-specified domain. Typical training corpora for LMs contain natural language (e.g. collected from the web), but LMs can also be trained on other types of languages (e.g. computer programming languages). Moreover, LMs can serve different purposes, such as generating language or providing semantic embeddings. Depending on the primary purpose of a LM, slightly different architectures and training objectives can be used. In this paper, unless we specify otherwise, we focus on LMs tailored to language generation.

A standard approach to construct generative LMs is to use an autoregressive decomposition that sequentially proposes a probability distribution for the next utterance based on past utterances:

\[
p(w) = p(w_1) \cdot p(w_2|w_1) \cdots p(w_T|w_1, \ldots, w_{T-1})
\]

Here \( w = w_1 \ldots w_T \) is a sequence of \( T = |w| \) utterances. Each of the terms \( p(w_t|w_1, \ldots, w_{t-1}) \) with \( t = 1, \ldots, T \) represents the probability the model assigns to observing the particular utterance \( w_t \) given the previous \( t - 1 \) utterances. LMs of this form are trained by updating the parameters controlling these conditional probabilities to assign high likelihood to sequences of utterances observed in the training corpus. Training is the result of an iterative process whereby at each iteration the model is presented with a batch of utterances and its parameters are updated to increase the likelihood of that particular set of utterances. Training large-scale language models can require very high numbers of iterations, requiring significant computing power.

Recent LMs are primarily distinguished from other LMs due to their parameter size and training data. Their size allows LMs to retain representations of extremely large text corpora, resulting in much more general sequence prediction systems than prior LMs.

In this report, we focus on such large-scale models.

Note that LMs do not output text directly. Rather, they produce a probability distribution over different utterances from which samples can be drawn. Greedy decoding directly from the (conditional) probability distribution provided by an LM is possible, but often performs poorly in practice. Instead, methods that focus on the most likely utterances – while introducing a small amount of variability (e.g. beam search and nucleus sampling) – have been found to produce better results in practice [78].

A.2 “Large” Language Models

Training models with a large number of parameters on extremely large datasets such as the Colossal Clean Crawl Corpus (C4) [149], WebText [147], and MassiveText [148] resulted in sequence prediction systems with much more general applicability compared to the prior state-of-the-art [30, 54, 148, 160] as well as impressive few-shot and zero-shot learning capabilities [30, 148]. The insight that powerful sequence prediction systems could be created by scaling up the size of LMs and training corpora motivated an upsurge in interest and investment in LM research by several AI research labs. The increase in size of LMs has implications for a number of risks discussed in this paper, for example an increase in quality of outputs in some tasks may increase the risk that users give credence to misinformation provided by the model in other tasks subsection 2.3; and all else held constant, environmental cost of LMs grow with their size.

A.3 Bias

Concerns regarding “bias” in language models generally revolve around distributional skews that result in unfavourable impacts for particular social groups [82, 169]. We note that there are different definitions of “bias” in classical statistics and machine learning compared to sociotechnical studies. In classical statistics, “bias” designates the difference between a model’s prediction and the ground truth [47]: in machine learning, minimising statistical bias is a component of reducing error [47]. In sociotechnical studies, “bias” refers to skews that lead to unjust discrimination based on traits such as age, gender, religion, ability status, whether or not these characteristics are legally protected [22]. Developing mechanisms to quantify the latter type of bias is an area of active research, where qualitative and quantitative measures have been established [12, 71].

A.4 Discrimination

Similarly, “discrimination” has a dual definition. Traditionally in machine learning, this term refers to making distinctions between possible categories or target classes [28]. In sociotechnical work, “discrimination” refers to unjust differential treatment, typically toward historically marginalised groups. Various steps in training a machine learning model can result in discrimination in the sociotechnical sense, from labelling and collection of the training data, to defining the “target variable” and class labels, to selecting features [13].
<table>
<thead>
<tr>
<th>Risk area</th>
<th>Mechanism</th>
<th>Type of Harm</th>
<th>Technical mitigation approaches</th>
<th>Evidence of this risk in large-scale LMs</th>
</tr>
</thead>
</table>
| **Discrimination, Hate Speech and Exclusion** | The LM accurately reflects unjust, toxic, and oppressive speech present in the training data. | • Allocational or representational harm  
• Profound offence or psychological harm  
• Inciting violence or hate  
• Social exclusion  
• Uneven performance for different social groups | • More representative training data  
• Curated and filtered training data  
• Dataset documentation  
• Participatory approaches for detecting instances of harm  
• Online learning for model updating  
• Training retriever model with separate data corpus  
• Prompt design  
• Explainability and interpretability research to identify fairness concerns | • Social Stereotypes: Abid et al. [2], Huang et al. [82], Lucy and Bumman [120], Nadem et al. [133], Nozza et al. [136], Rae et al. [148]  
• Hate Speech: Gehman et al. [61], Rae et al. [148], Welbl et al. [194]  
• Exclusion: [206]  
• Lower Performance: Welbl et al. [194], Winata et al. [196] |
| **Information Hazards** | The LM leaks or correctly infers sensitive information. | • Privacy violations  
• Safety risks | • Algorithmic tools such as differential privacy  
• Responsible release strategies | • Privacy leaks: Carlini et al. [34], Kulkarni [109] |
| **Misinformation Harms** | The LM provides false, misleading, nonsensical or poor quality information. | • Deceiving or misinforming a user  
• Material harm  
• Unethical actions by users  
• Growing societal distrust in shared information | • Responsible release strategy  
• Innovate on methods to filter out incorrect statements  
• Training retriever model with separate data corpus  
• Engineer LMs to not provide output on sensitive domains  
• Sociotechnical interventions such as training truthfulness via humans-in-the-loop, and shaping norms and institutions on truth in the field  
• Training LMs that can search and reference sources from the internet to substantiate factual statements | • Misinformation: Hendrycks et al. [72], Lin et al. [118], Nokano et al. [134], Quach [145], Rae et al. [148], Reiter [157] |
| **Malicious Uses** | Humans intentionally use the LM to cause harm. | • Undermining public discourse  
• Facilitating fraud, scam, impersonation crimes  
• Personalised disinformation campaigns  
• Weaponisation or production of malicious code  
• Augment illegitimate mass surveillance | • Limit access to the LMs and monitoring usage | • Disinformation: Bagdasaryan and Shmatikov [11], Buchanan et al. [31], Hao [69], Zellers et al. [203] |
| **Human-Computer Interaction Harms** | Humans are deceived or made vulnerable via direct interaction with a powerful conversational agent. | • Unsafe use  
• Creating avenues to exploit or violate privacy of the user  
• Perpetuating discriminatory stereotypes via product design | • More inclusive product design  
• Giving an assistant non-gendered or multiple voices | • Product design: Cave and Dihal [35], Zdenek [202]  
• Exploiting users: Ischen et al. [87], Parides [139]  
• Cybersecurity Chen et al. [37] |
| **Environmental and Socioeconomic harms** | LMs are used to underpin widely used downstream applications that disproportionately benefit and harm different groups. | • Increasing social inequalities from uneven distribution of risk and benefits  
• Loss of high-quality and safe employment  
• Undermining creative industries  
• Environmental harm | • Architectural innovations such as training retriever model with separate data corpus  
• Increase training efficiency | • Environment: Bender et al. [15], Myttiton [132], Patterson et al. [141], Rae et al. [148], Strubell et al. [176]  
• Unequally distributed benefit: Bender et al. [15] |