Why academia should cut back on general enthusiasm about CAs

Alessia Giulimondi
Utrecht University
a.giulimondi@students.uu.nl

Abstract

This position paper will analyze LLMs, the core technology of CAs, from a socio-technical and linguistic perspective in order to argue for a limitation of its use in academia, which should be reflected in a more cautious adoption of CAs in private spaces.

The article describes how machine learning technologies like LLMs are inserted into a more general process of platformization, negatively affecting autonomy of research. Moreover, fine-tuning practices, as means to polish language models are questioned, explaining how these foster a deterministic approach to language.

A leading role of universities in this general gain of awareness is strongly advocated, as institutions that support transparent and open science, in order to foster and protect democratic values in our societies.

1 Introduction

Ever since the deployment to the general public of Chat-GPT, the public debate has promptly polarized around an excessive enthusiasm or an equally disproportioned alarm toward systems like Conversational Agents (hereafter CAs), that employ a technology supported either partially or entirely by machine learning algorithms. Recognizing a frequent failure in identifying the strong continuity these technologies manifest with their predecessors, this position paper will try to connect to each other the very diversified criticalities of the deployment of these technologies to identify the risks they pose for individual and institutional autonomy.

This position paper will focus on a specific class of deep learning architectures, Large Language Models (LLM), since they represent the core technology that constitutes Conversational Agents (CAs). This paper will advocate that specifically academia, given its (in principle) neutral position in society, in between and above interests of the market and the necessities of the state and its government, should take a clear stand for the limitation of LLMs, as industry-driven technologies, in its domain of competence (i.e. universities, research institutes). By doing so, academia may be able to influence society regarding its judgement and attituted toward devices such as CA that in most cases rely on LLMs.

Relying on LLMs, CA belong to those technologies developed and deployed by American high-tech private companies and growing literature (Van Dijck et al., 2023; Van Dijck, 2021; Benn and Lazar, 2022; Tafani, 2022; Hovy and Spruit, 2016) is showing the serious ethical and socio-technical problematics of a technological development that increasingly witnesses a concentration of power in a few private enterprises. Moreover, various comprehensive analyses have highlighted the negative effects of LLMs and data collection and profiling practices more in general (Couldry and Mejias, 2019; Weidinger et al., 2022; Matz et al., 2023; Andrić and Kasirzadeh, 2023). However, the alarming implications of these studies for individual and institutional autonomy, despite authors do not abstain to account for them, are too often underestimated.

It is sensible to argue that this general attitude to consider of secondary importance the warnings of scholars, technicians and activists (Chomsky, 2023; Harari, 2018) about the dangers of a techno-economic monopoly and user profiling undermines the soundness of technological development itself, as well as the integrity and autonomy of public research institutions (Kersessens and van Dijck, 2022). Indeed, autonomy of universities is crucial to ensure that democratic values are preserved in a society. Nevertheless, critiques to the current modalities of technological development are often casted aside, mostly perceived as a major impediment to the progress of society (Tafani, 2022) – arguably implying an understanding of progress only driven and substantiated by technical advance-

Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024) , pages 9–15
March 22, 2024 ©2024 Association for Computational Linguistics

ment.

Universities should foster and protect free, autonomous and creative research as a foundational value of its existence, resulting in a natural opposition to any policy, favored by the implementation of specific designs of a technology, that hinders in any way the free practice of autonomous scientific inquiry. Indeed, it is important to recognize how philosophical reflections, promoted by autonomous academic research, must sustain technological development, by ensuring a solid base for ethical decisions that (should) have the ultimate say in the deployment to the society at large of technically complex tools intended to serve and support (when not replace) very diverse human activities.

This article will describe the dependency of LLMs from a platform society (2.1) to argue for a lack of autonomy of research in the presence of a techno-economic monopoly. Section 2.2 will describe how this results in a structural dependency of academia from big tech companies and section 2.3 will argue for the inappropriateness of the use of opaque technologies in academic research. In section 3 the processes involved in fine-tuning will be outlined to question the validity of models that propose a deterministic and one-sided view of language. Thus, a critical analysis of the implications of leaving to private companies the modeling of language and the use of these models to generate language in robots or devices (e.g. CAs) is proposed. By evaluating the critical characteristics of CAs through the examination of the problematics of LLMs, this paper will advocate for the necessity of academia to assume a leading role in the identification of the dangers of the dominating trends of technological development, to favor a shift toward more democratic ones.

2 LLM and CAs as an opaque product of a platform society

2.1 Platformization

Research conducted by José van Dijck in the last decade is of fundamental importance to gain a complete picture of the socio-technical context in which LLMs and CAs are developed.

What van Dijck proposes is an analysis of the recent socio-political and socio-economic changes within the frame of a platformization process, a process that is transforming our societies as industrialization did in the past. Platformization is described as a dynamic that happens within a platform ecosystem, understood as “a corporate space” (Van Dijck et al., 2023, p .3441) that is commonly known to be controlled by five American private companies (Google, Apple, Facebook, Amazon, Microsoft, GAFAM). This ecosystem is a socio-technical infrastructure that is able to penetrate the public and private spheres, increasing the dependence of the latter on their services. Van Dijck described the complex dynamics that govern this platform ecosystem, “hierarchical and proprietary in nature” (Van Dijck et al., 2023, p .3441) as a data-driven system which survives nourished by a continuous collection of data. This ecosystem presents a layered structure of three levels where the Big Five have been increasing their presence and control. The deeper level of the infrastructure is constituted by underwater cables and data centers that ensure the collection and distribution of data, while the intermediate level includes the cloud services necessary to process the data. The sectoral applications (e.g. mobility or educational apps) depend on these “lower” infrastructures and the vertical integration across the three levels of infrastructures. Increasing the dominance of private corporations in the deeper level and the intermediate level is resulting in an overall privatization of the Internet space which was initially intended to overcome the geopolitical barriers and interests to serve as a “utility, independently organized and managed” (Van Dijck et al., 2023, p.2805).

Therefore, proposing an analysis that considers the advent of Artificial Intelligence in strict continuity with this progressive and accelerated penetration of private infrastructures into spheres that traditionally belong to the public domain is perhaps not ventured. Indeed, van Dijck points out how a specific feature of this ecosystem is its capability of posing itself outside of the traditional limits of the public and private spheres. In other words, this companies built a system which survives on the exceptionality of their position within the civil society, making laws and regulations hard to be applied. “Tech companies deliberately push their platforms to vacillate between sectors and infrastructures, between markets and nonmarkets, between private and public interests, between a marketplace for goods and services and a marketplace of ideas, while adopting features of both.” (Van Dijck, 2021, p. 2810)

2.2 A structural dependency

The deployment to the public of Chat-GPT falls under this well-established process of releasing technologies that are nurtured in their roots by the monopolizing nature of the system that created them (Van Dijck, 2021). This ecosystem sustains itself by datafication, strictly connected and dependent on the platformization described in 2.1. Datafication is the process of transforming activities performed online into data-points exploitable by private and/or public companies and institutions, and it is the result of the seamless flow of data across the three layers of the system, which ensures a solid control of the entire infrastructure that supports data collection and distribution (Couldry and Mejias, 2019). Eventually, this made possible for tech companies to gather data deep enough to train a model that has enough parameters (or weights) to perform surprisingly well in NLP and language generation, paving the way for more human-like CAs (Couldry and Mejias, 2019). However, it is important to mention studies that have recently shown how it is possible to have seemingly performant language models that rely on smaller datasets (van Dijk et al., 2023). Thus, it is perhaps possible to imagine artificial neural networks that do not necessarily rely on immense quantity of human data. Nevertheless, it is generally acknowledged the monopoly of the Big Five in the development and deployment of LLMs, which naturally draws the attention of researchers concerned about the ethical implications of these models to the ones that are most commonly used in both private and public contexts. The benignity of some language models trained independently, on non-opaque datasets with transparent methodologies in the pre-training and fine-tuning phases do not fall in the scope of this paper. It is the monopoly of tech-companies over LLMs and the consequent imposition of their theoretical assumptions and designs what poses serious concerns for the autonomy of research, since in the majority of cases universities have to rely at least on the pre-training phases provided by private tech companies (Kersessens and van Dijck, 2022). The power gained in the last decades by GAFAM across the layers of the digital infrastructure makes the creation of an independent system extremely costly (Karpathy, 2023) and ostensibly less efficient for any small tech enterprise whether private or funded by the university or the government.

This is a first direct influence of private corporations over public educational institutions, such as the university. Through the appropriation of expertise and infrastructures, they offer researchers (as well as private users) the only choice of selecting one of the few companies that are able to provide highly performing digital services (from software and cloud services to language models like Chat-GPT), inevitably "imposing their architecture choice design upon users" (Van Dijck, 2021, p. 2810). This architecture choice becomes an imposition in absence of a fair market in which a truly diversified range of possibilities is offered to users and institutions. Moreover, it becomes an even more unsettling scenario when the freedom itself for institutions to build their own platforms, that abide to the rules decided within that institution, is heavily limited by the privatization of the infrastructures of the Internet. Therefore, a first direct impact on autonomy of research is arguably observable when a specific architecture for LLMs is deployed and researchers are urged to adopt them, often mostly on the basis of their performativity. It is possible to counterargument that it is in the very nature of research employing the best tools available on the market for ends that might be beneficial for research itself and for society at large. However, this view heavily undermines the freedom of scientific inquiry that, in principle, should not become entirely dependent on companies primarily driven by interest of profit, in order not to erode the fundamental difference between academic research and industrial research (Kersessens and van Dijck, 2022).

2.3 LLMs: an opaque technology

A second aspect that should be taken into consideration when talking about CAs and their use in both private and public spheres is that LLMs, their core technologies, rely on data collection and processing that are notably opaque in their nature (Vetrò et al., 2019; Couldry and Mejias, 2019; Tafani, 2022; Andrić and Kasirzadeh, 2023). The creation of parameters (or weights) during the training stage of these larger models (e.g. Chat- GPT, Llama, Gemini) is arguably one of the most controversial part, as this is the stage where researchers admit the lowest level of control over the process. Andrej Karphaty, in one of his lectures, explains clearly how artificial neural networks are treated by computer scientists involved in their development as "mostly inscrutable, empirical artifacts" (Karpathy, 2023). Therefore, a structural opaqueness lies at the core of larger LLMs and this already poses some concerns about the appropriateness of employing such artifacts as tools intended to support academic research, while valuing transparency as a fundamental principle for a more open science. A straightforward example of the negative repercussions of this opacity (together with a frequency-driven nature of the model) is that "LMs are trained to predict the *likelihood* of utterances", which does not predict its correctness and "this may present a theoretical limit on LM capabilities to detect misinformation" (Weidinger et al., 2022, p. 218). Furthermore, it was shown how this opaqueness does not conform to privacy regulations and democratic principles that constitute the foundations of substantial freedom in democratic societies (Weidinger et al., 2022; Andrić and Kasirzadeh, 2023; Couldry and Mejias, 2019). Thus, universities and research institutes that support democratic values within a society have the social responsibility to limit, at least within its direct domain of actions, the indiscriminate adoption of highly controversial technologies on the socio-technical and socio-economic level.

It is also true, however, how a first, crucial, step toward this desirable academic policy is to recognize the controversial status of these technologies. Nevertheless, due to the blurred distinction between public and private sectors created within this platform ecosystem (see 2.1), it is often not easy also for researchers to spot the ambiguities and criticalities of these processes, as university and research more in general are themselves part of this process of platformization (Kersessens and van Dijck, 2022). Indeed, the lack of awareness of university's overreliance on Big Tech companies' infrastructures and services appears to be a rather established phenomenon, as it is demonstrated by the regular practice to use Google Scholar as a starting point of any literature search, which lies at the foundations of any scientific investigation. The implications for diversity of the literature and the consequent autonomy of scientific studies are rarely sounded out.

Thus, to conclude this first section, it is reasonable to view LLMs and CAs as a general phenomenon of strengthening the dominance of big tech companies in both public and private sectors (including academia) and a natural continuation of the development of technologies that mostly follow the logics of the market. In the following section, we will analyze how the socio-political problematics highlighted in this first part of the article cannot be completely separated from the more technical concerns that can be raised regarding the development of language models, that follow the structures and logics of a commercial company.

3 Modeling language for machine learning: is it really appropriate?

In linguistic research, the practice of creating language models is uncontroversial and largely employed to propose and explain theories of language. Models enable us to visualize and better understand the mechanisms of phenomena like language that are not directly accessible via simple observations and descriptions. We need theories and hypotheses to model language and we need models to argue for those same theories and hypotheses. A good definition of a scientific model is the one that defines it as a "visualization of entities non representable in other ways, in their reduction to an empirical description, in the simulation of the logico-structural characteristics of a research object, via the creation of isomorphisms and analogies." ('modello', 2023). van Dijk et al.'s (2023), explains the large potential these language models have for research in language acquisition, as they represent a statistical model that informs us of the possibilities of statistical learning likely at play in language acquisition. Set aside the ethical controversies connected to the use of opaque systems, this can be an overall correct use of a language model. Indeed, in this type of research scenario, the language model would serve the role of a model that supports the scientific understanding of reality (language acquisition, in this case). On the other hand, when (large) Language Models are fed into a system created specifically to interact with humans, like in the case of CAs, the situation substantially changes. Firstly, the model, initially intended as a hypothetical approximation of how language works, becomes a generative system that is meant to imitate a natural, human phenomenon like speaking. More concretely, CAs are intended to use a language that meets syntactic, pragmatic and discourse standards that are inevitably decided by their developers (Karpathy, 2023; Kasirzadeh and Gabriel, 2023). These standards are manually inserted during the fine-tuning process, which needs human intervention to categorize responses that are considered appropriate or correct. The problem of this common practice, known as Reinforcement Learning from Human Feedback (RLHF), is the aprioristic choice that lies beneath any categorization of this kind. Indeed, LLMs undergo two phases of their training: the first one (pre-training) where parameters are created and fed into the neural network and the second phase, which prepares the model to be able to answer questions. This second phase consists of a process called fine-tuning. Fine-tuning makes use of labels that have to be assigned to different types of potential responses a CA can give to the user. This is meant to make a first alignment to human conversational conventions, by teaching the algorithm which responses are more desirable or more correct. Fine-tuning is commonly the phase where researchers try to operate most interventions (Kasirzadeh and Gabriel, 2023) to reduce toxicity, inappropriateness and biases of the model often shown to be a major issue for social discrimination and perpetuation of stereotypes (Weidinger et al., 2022; Andrić and Kasirzadeh, 2023). However, the question of whether it is really possible to clean these models from biases and what this really entails is often avoided. In a recent talk held at the Symposium of philosophy of science, AI and machine learning, Tom Sterkenburg described how biases are rather natural outcomes of LLMs because of the naturally biased nature of human data on which they are trained (Sterkenburg, 2023). Moreover, he also explained how this biases are also model-dependent. Thus, it is perhaps a “false problem” to talk about biases in LLMs, and focusing the large part of AI ethic research on the removal of biases that cannot ultimately be removed risks to be counterproductive. Nevertheless, what it seems even more crucial in order to understand the need for a change in perspective is to further analyze the implications of this “fight” against biases of the algorithms. Kasirzadeh and Gabriel’s (2023) proposed an application of the knowledge of pragmatics to CAs, employing Gricean maxims and Speech Act theory to elaborate a set of rules that an ideal CA should follow in order to be a desirable conversational partner. The elaboration of these rules are intended to propose a pragmatic approach for the long-standing problem of what can be considered a human value general enough to be universally extended. However, the proposal only succeeds in demonstrating the methodological inadequacy of applying linguistic theory to commercial products meant to be used by a wide variety of people in a large set of diversified contexts. Linguistic theories, such as Speech Act theory and Grice maxims (Huang, 2016; Mabaquiao, 2018) do not have to be understood as rules humans have to follow to have a successful conversation. Rather, they were proposed to describe conventions and general patterns that are hypothesized to be at play in human linguistic interactions. Thus, they should not be understood as directly applicable to automated processes. Indeed, the essential problematic of CAs is the automation that lies at the core of its functioning. Automating a process such as language, which linguists still struggle to understand as a phenomenon and which manifests itself as a creative, continuously changing and evolving process is intrinsically problematic. Automating it on the base of a model mostly grounded on statistical probabilities and a subsequent labeling process may easily lead to a deterministic view of language, with non-negligible consequences for autonomy of the user repeatedly exposed to a pre-determined language. This approach is in line with a more general approach, often referred to as a "new behaviorism" (Tafani, 2022). In this regard, we signal the thorough analysis conducted by Benn and Lazar’s (2022) about Automated Influence.

Therefore, the problem does not resolve itself on the decision of which type of labels is best to assign, but the concerns lay on the very nature of the labeling process necessary for fine-tuning. Indeed, labeling excerpts of texts can endanger freedom of thought and expression, as it implicitly conveys what is allowed and what is better avoid saying. This can be argued from evidence we have from speech alignment (Pickering and Garrod, 2004) that showed how interlocutors tend to align to the language of their addressee on various linguistic levels (syntactic, lexical, phonological, pragmatic). Thus, it is reasonable to hypothesize that a similar pattern of alignment can occur also when the interlocutor is not human, but it successfully imitates human language. The intention here is not to argue for a direct impact of CAs on users, as in an online alignment to the CA which eventually results in a permanent alteration of language use of the individual user. There is no scientific basis to hypothesize such an outcome. The purpose of this last consideration is rather highlighting how speech alignment studies can inform us about the capacity language has to shape and modify itself and its environment according to necessities and contexts, and how this is directly linked to how humans adapt and adjust depending on the situations and interlocutors. Thus, it is important to critically understand what it means to engage in various conversations with devices that successfully resemble human language, while this resemblance is a product of very different mechanisms from the ones that operate in the human brain (van Dijk et al., 2023). Indeed, one of the inherent characteristics of human language is precisely the capacity of creating an infinite set of possible outputs given a finite set of items (Hauser et al., 2002). Setting aside the fact that the state of LLMs seems to resemble quite the opposite situation, the free creation of linguistic material, highly interrelated with thought generation and its free expression, can be seriously challenged by a view that considers language a large set of items that can be labeled according to pragmatic conventions, wrongly interpreted as rules to follow, and policies that set standards for what it can be considered civil to say. Thus, more research is needed that addresses this issue, *before* deploying to the public technologies of which the long-terms effects are mostly unknown.

It is now, perhaps, easier to understand why the problems with language modeling are also deeply interrelated with the fact that LLMs that are more commonly used are mostly developed by private corporations that are inevitably imposing a unilateral, English-centered and Western-centered model of language – for obvious reasons connected to the centralization of tech-power in American companies described above. Therefore, it is not easy to disentangle the problems inherent to language models and problems related to the monopoly over these by GAFAM. In other words, it is difficult to research LLMs and their design with sufficient objectivity, if what is currently mostly available on the market is only one way of doing things (with few variations).

Freeing machine learning technologies from a monopoly that interests a large portion of the globe is crucial to ensure enough diversity in technological research and development, which is at the foundation of effective and meaningful research intended to benefit society as a whole. Researchers in the field of linguistics, data and computer science, electronic engineering, philosophers, psychologists and sociologists should be able to conduct their individual and collaborative work independently, both from a socio-psychological perspective and technical perspective. Indeed, they should not be dependent on private companies for the delivery of tools and expertise, nor they should suffer from an imposition of a specific design of language and interaction, mostly designed to induce the user to stay hooked to the device (Matz et al., 2023; Hovy and Spruit, 2016; Couldry and Mejias, 2019).

In order for societies and governments to envision these problematics and promote practices and regulations that support an open and democratic technological advancement, it is essential that the university, as a social party that historically fosters diversity of thought and free creation and circulation of knowledge, takes a clear stand in the limitation of the expansion of highly controversial technologies in research and society.

4 Limitations and conclusions

As it was already pointed out by van Dijk et al.'s (2023), deep learning technologies are a "moving target" considering the fast pace at which their training and deployment is moving. For this reason, it is not possible to discuss the abilities LLMs will have in some months. However, this state of the art should prompt academia to reflect once again on the appropriateness and overall safety of this general acceleration, driven by unconstrained release of technologies by private companies that puts researchers in serious difficulties when attempting to investigate with lucidity and transparency these tools.

Moreover, it is not among the intentions of this article to deny the numerous benefits LLMs may have for research. Indeed, the implicit proposal of this dissertation is the distinction between LLMs used for research purposes – once ensured autonomy, fairness and transparency – and LLMs implemented in CAs, meant to be used by private users for different goals. Within this distinction, benefits for research are positively reviewed, while benefits for the private user are questioned. Furthermore, the fine-tuning practices largely employed with an intention to improve the model and polish it from biases and toxicity are critically reviewed. It is, indeed, proposed a view that questions the validity of automation of language from a methodological perspective, arguing that it supports a deterministic approach to linguistic data and human behavior more in general.

Finally, we limit our critique to LLMs that employ parameters and architectures financed, developed and/or supported by private corporations that hold a great asymmetric power with public institutions and governments across the globe. Whether this interests the vast majority of LLMs currently used, it is left to the judgement of the reader.

References

Katja Andrić and Atoosa Kasirzadeh. 2023. Reconciling Governmental Use of Online Targeting With Democracy. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1871–1881, New York, NY, USA. Association for Computing Machinery.

Claire Benn and Seth Lazar. 2022. What's Wrong with Automated Influence. *Canadian Journal of Philosophy*, 52(1):125–148. Publisher: Cambridge University Press.

Noam Chomsky. 2023. Opinion | Noam Chomsky: The False Promise of ChatGPT.

Nick Couldry and Ulises A. Mejias. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, 20(4):336–349. Publisher: SAGE Publications.

Yuval Noah Harari. 2018. Why Technology Favors Tyranny. *The Atlantic*. Section: Technology.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? In Hiroko Yamakido, Richard K. Larson, and Viviane Déprez, editors, *The Evolution of Human Language: Biolinguistic Perspectives*, Approaches to the Evolution of Language, pages 14–42. Cambridge University Press, Cambridge.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Yan Huang. 2016. Conversational Implicature.

Andrej Karpathy. 2023. [1hr Talk] Intro to Large Language Models [Video].

Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2):27.

Neils Kersessens and José van Dijck. 2022. Governed by Edtech? Valuing Pedagogical Autonomy in a Platform Society. *Harvard Educational Review*, 92(2):284–303.

Napoleon Jr Mabaquiao. 2018. Speech act theory: From austin to searle. 19:35–45.

Sandra Matz, Jake Teeny, Sumer S. Vaid, Gabriella M. Harari, and Moran Cerf. 2023. The Potential of Generative AI for Personalized Persuasion at Scale. Publisher: OSF.

'modello'. 2023. Treccani.

Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–225. Publisher: Cambridge University Press.

Tom Sterkenburg. 2023. Epistemology and theory of machine learning.

Daniela Tafani. 2022. What's wrong with "AI ethics" narratives. Publisher: Zenodo Version Number: Published.

José Van Dijck. 2021. Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, 23(9):2801–2819. Publisher: SAGE Publications.

José Van Dijck, Tim de Winkel, and Mirko Tobias Schäfer. 2023. Deplatformization and the governance of the platform ecosystem. *New Media & Society*, 25(12):3438–3454. Publisher: SAGE Publications.

Bram M. A. van Dijk, Tom Kouwenhoven, Marco R. Spruit, and Max J. van Duijn. 2023. Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. ArXiv:2310.19671 [cs].

Antonio Vetrò, Antonio Santangelo, Elena Beretta, and Juan Carlos De Martin. 2019. AI: from rational agents to socially responsible agents. *Digital Policy, Regulation and Governance*, 21(3):291–304. Publisher: Emerald Publishing Limited.

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 *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea. ACM.