JDSR 7(1) (2025) 66–82

JOURNAL OF DIGITAL SOCIAL RESEARCH

10.33621/jdsr.v7i154879

Autocompleting inequality

Large language models and the "alignment problem"

Mike Zajko

University of British Columbia Okanagan, Canada

✉ mike.zajko@ubc.ca

Abstract

The latest wave of AI hype has been driven by 'generative AI' systems exemplified by ChatGPT, which was created by OpenAI's 'fine-tuning' of a large language model (LLM). This process involves using human labor to provide feedback on generative outputs in order to bring these into greater 'alignment' with 'safety'. This article analyzes the fine-tuning of generative AI as a process of social ordering, beginning with the encoding of cultural dispositions into LLMs, their containment and redirection into vectors of 'safety', and the subsequent challenge of these 'guard rails' by users. Fine-tuning becomes a means by which some social hierarchies are reproduced, reshaped, and flattened. By analyzing documentation provided by generative AI developers, I show how fine-tuning makes use of human judgement to reshape the algorithmic reproduction of inequality, while also arguing that the most important values driving AI alignment are commercial imperatives and aligning with political economy.

Keywords: generative AI; alignment; inequality; language

1. Introduction

In early February 2023, numerous news outlets and politically conservative voices shared versions of a story in which OpenAI's popular chatbot, ChatGPT, refused to condone the use of any racial slur, even in a ridiculous scenario where racist language could somehow save millions of lives (Aleem, 2023). This was one of several instances of conservative backlash against apparently progressive (or "woke") values being reproduced by chatbots (Tiku & Oremus, 2023). In a more recent example, Google's Gemini AI system was widely criticized for "inaccurate" depictions of historical characters, demonstrating what many saw as an excess of gender and racial diversity (Edwards, 2024).

All of this is a markedly different dynamic than that found in earlier sociological critiques of AI (see Benjamin, 2019; Joyce et al., 2021), wherein algorithmic technologies reproduce racism, sexism, and more nuanced forms of inequality and 'bias'. More than a decade ago, Google was criticized for providing users with racist and sexist autocomplete suggestions and search results, thereby reinforcing oppressive social relations (Noble, 2018). In response to media attention, Google explained that these results were based on users' behavior and interests, but did take steps to remove them (Auerbach, 2013; Gibbs, 2016). The corporate risks of chatbots powered by language models were most clearly demonstrated in 2016, when Microsoft had to withdraw its Tay chatbot after users (the "trolls" of 4chan) found how to shift its propensity to produce racist and sexist outputs (Schwartz, 2019). In subsequent years Tay was followed by other examples of chatbots “going off the rails” (Hao, 2023), such as Lee Luda, the South Korean chatbot that had to be shut down amid scandal in 2021 (McCurry, 2021).

To avoid similar controversies, major generative AI developers have ‘aligned’ their chatbots towards non-discrimination. When asked to comment on marginalized groups, these services typically affirm fundamental human equalities and push back against derogatory language. Instead of “racist robots” (Benjamin, 2019), today’s generative AI algorithms are avowedly anti-racist on the surface, despite the racism in the hidden layers shaped by their training data. The ‘guard rails’ separating the two are the result of ‘fine-tuning’ by workers hired to pass judgement on the model’s language use, and this becomes a key site of social ordering and iterative social struggle. Social inequalities are introduced and reproduced through training data, partially neutralized through human feedback and guard rails, then resurfaced through red-teaming and jailbreaking, and neutralized again in a recurring fashion.

The metaphors of fine-tuning, guard rails, and resurfacing remind us that these are largely superficial struggles over social inequality, rather than deeper, structural changes. At issue are the public-facing outputs of generative AI systems, and the corporate investments in ensuring that these outputs are equality-affirming have been driven by concerns over the ‘reputational risk’ of being associated with offensive language. However, struggles over how human groups are represented do have major consequences for human lives, and these are magnified as LLMs become more widely-deployed in various uses.

In this article, social inequalities are understood as asymmetric forms of group differentiation that contradict a normative positioning of these groups as equals. This “traditional” conception of inequality (as “a mathematical-normative hybrid”) “implies injustice” (Hirschauer, 2023, p. 362), in that it concerns differences that are illegitimate or in opposition to fundamental democratic equalities. Social inequality becomes a social problem in a political context that is organized around affirming equalities between particular human categories, such as gender and race (see Rosanvallon, 2013). Blatant inequalities are also a business problem for new commercial services that seek legitimacy and to avoid scandal. While the guard rails of generative AI are justified as the pursuit of ‘safety’, they are primarily intended to protect the commercial viability of generative AI systems.

To theorize the relationship between AI and social inequality, this research builds on a Bourdieusian perspective that has been valuable in connecting the cultural reproduction of social order with machine learning (Airoldi, 2022; Fourcade & Johns, 2020). While this approach has been useful in explaining how existing hierarchies are reproduced through AI, of primary interest here is an explanation of how generative AI has been ‘tuned’ to avoid reproducing particular inequalities (namely sexism and racism). Doing so requires attending to how the work of fine-tuning is textually mediated and coordinated towards certain goals across time and space. However, to understand what the goals or ‘values’ of fine-tuning are, requires grounding our analysis in political economy. This is because generative AI has been an expensive investment in what is intended as a profit-making enterprise. Commercial exploitation is a primary consideration in “data work” (see Miceli & Posada, 2022; Miceli, Schuessler & Yang, 2020), and the cultural reproduction of other forms of oppression can actually be a threat to business interests. Therefore, my argument is that AI’s alignment problem is not about “aligning with human values” (Askell et al., 2021) in terms of what humans might broadly want from AI systems, but is instead a problem of aligning these systems with political economy and whatever is conducive to commercialization. To the extent that these systems are being aligned towards equality, this remains a particular (liberal) form of equality oriented towards equal treatment or neutrality, particularly along lines of gender and race, rather than more radical or transformative alternatives. The efforts of these commercial actors provide a valuable demonstration of the possibilities and limits of shifting inequalities in code, which can be pursued with greater ethical care towards other ends.

2. Methods

Studying fine-tuning in generative AI as a social process is a challenge given the multiple stages of development and actors involved, which typically operate under a shroud of corporate secrecy. The analysis that follows draws on a variety of published materials, but is based in large part on documents made available by three generative AI developers (OpenAI, Anthropic, and Meta) about their fine-tuning processes. This includes Anthropic's (2022) human feedback and red-teaming datasets, which contain tens of thousands of interactions between chatbots and the data workers tasked with fine-tuning their responses. The articles (Bai et al., 2022; Ganguli et al., 2022) published by Anthropic about this work provide the instructions used to guide this labour. Documentation from OpenAI includes the instructions used to fine-tune InstructGPT (Ouyang et al., 2022), a precursor to ChatGPT that has informed the company's subsequent work. As shown by Miceli and Posada (2022), instructions function as key texts in the hierarchical workplace relations that data workers are subject to, providing "predefined truth values" (p. 29) that can be consistently applied through data labelling.

While the most significant developers of generative AI (including OpenAI and Anthropic) have become quite secretive about their development process since 2022, the release of new generative AI models has sometimes been accompanied by documentation that provides some methodological details and fine-tuning examples. Specifically, I also analyze the "system cards" that accompanied OpenAI's release of GPT-4 (OpenAI, 2023b) and DALL-E 3 (OpenAI, 2023a), as well as a report from Meta accompanying the release of Llama 2 (Touvron et al., 2023). While these sources do not provide a complete set of fine-tuning instructions or comprehensive record of work as with the sources above, they do describe these companies' priorities and procedures for fine-tuning, illustrated with selected examples. Within these documents, I focus on aspects related to social inequality or differential treatment of human groups, attending to how certain kinds of inequality (namely gender and race) are prioritized for fine-tuning, and situate these within the larger discourse of 'safety' that has become the predominant way of discussing a wide range of undesirable behavior by generative AI. In other words, my analysis of the corporate documentation made available about fine-tuning attends to how these companies operationalize their concerns about inequality through specific 'mitigation' techniques, and how these efforts are discursively justified.

Finally, this article also draws some of my own experiences (in 2023-24) using generative AI services and experimenting with prompts – generating written narratives and images to examine tendencies in how people are represented. Systems such as ChatGPT have been repeatedly updated and outputs may vary each time they are generated, so these refer to tendencies observed on a particular date, as detailed in footnotes. While some of this work has been systematic, repeatedly regenerating outputs for prompts that can be compared with others, this analysis can be considered an "algorithmic poke" (Gillespie, 2024, p. 3) at best, rather than an algorithmic audit. There remains a need for scholars to more systematically document variations in chatbot responses and how these change or are updated over time.

3. Language and the reproduction of inequality

Over the past decade, critical scholarship has exposed various ways that algorithmic systems perpetuate inequalities (Benjamin, 2019; Eubanks, 2018; Joyce et al., 2021; Noble, 2018; O'Neil, 2016), but these are always in relation to pre-existing systems of stratification or social structures. Within these, language is a key means for the reproduction of hierarchies, as most famously theorized by Pierre Bourdieu, who wrote about how "linguistic capital" and "linguistic habitus" favor some individuals and groups over others, depending on what kinds of language are considered legitimate, authoritative, or vulgar (Bourdieu, 1991). An LLM also does not treat all language as equal, as determined by what is included and excluded in its training data, or how language is classified by its filters. Many datasets remain English-centric, and appear to favor values specific to the U.S. (Johnson et al., 2022). Inequalities exist among English speakers as well – a recent study showed that speakers of African-American English were more likely to be judged negatively by LLMs in terms of personal characteristics, criminality, and associated occupations (Hofmann et al., 2024).

In addition to the fact that the norms encoded in LLMs privilege and exclude different linguistic groups, there are also ways that speech, language, and discourse function to order and stratify the world, through the exercise of what Bourdieu (1991) sometimes characterized as "symbolic violence" (Airoldi, 2022, pp. 114–15). Because language is used to define social hierarchies, LLMs replicate this behavior and perpetuate language-based harms against a wide range of marginalized or stigmatized groups (Gallegos et al., 2024; Mei, Fereidooni & Caliskan, 2023). LLMs can be used to predict or "auto-complete" (Huang, 2023) text-based responses to human 'prompts', and when completing statements about various already-disadvantaged groups, they are more likely to do so with negative and disparaging language (Sabbaghi, Wolfe, & Caliskan, 2023), reinforcing negative outcomes for those groups.

Representational harms that have been studied in language models include the erasure of certain kinds of people from representation, the reification of essential differences between human categories, and the stereotyping of social groups (Shelby et al., 2023, pp. 728-29). A well-known example involves having a chatbot assign men and women in a gendered occupational hierarchy (Ghosh & Caliskan, 2023). The resulting output will routinely place a man in the superior position (ie. doctor, CEO) over a woman (ie. nurse, administrative assistant). Stories written by today's most popular chatbots tend to reinforce normative assumptions and identities, such as heteronormativity (Gillespie, 2024), marginalizing representations of other kinds of people and relationships.

The automated reproduction of inequality in generative AI can be conceptualized in broadly Bourdieusian terms as "machine habitus": encoded cultural dispositions as statistical propensities in a computer model, allowing for the "conscienceless reproduction of recurrent data patterns" into new cultural products (Airoldi, 2022, p. 60). It is important to reiterate that these patterns are derived from statistical propensities in the model's training data, rather than actual distributions of human characteristics, tendencies, or social divisions. Hence, we see a "Muslim-violence bias" from LLMs trained largely using English-language content scraped from websites (Abid, Farooqi, & Zou, 2021), while image generators are predisposed to sexualizing women or girls and whitening their features (OpenAI, 2023a; Snow, 2022), due to a large portion of the training data consisting of sexualized photos of light-skinned women. Key features of existing social hierarchies may be 'mirrored' in model outputs, such as the tendency for white men to occupy positions of power (Jacobi & Sag, 2024), but model outputs gravitate towards averages in the training data that can actually translate into less diversity than exists in the world.

While the reproduction of gender stereotypes in language reproduces or amplifies social hierarchies, Gross (2023) argues that generative AI can be a site of social change or a means to "undo gender" (see also Fournier-Tombs, 2023). This might mean making gender irrelevant in chatbot responses, or actively counteracting gendered biases and stereotypes. This optimistic possibility is premised on the fact that while generative AI systems require a great deal of labor time and capital to train, they can also be re-trained or fine-tuned with other priorities in mind. An update to a single, widely deployed AI system can have widespread consequences for social inequality; language and values can be reconfigured to propagate through AI outputs and shape society accordingly.

My argument is that generative AI has already become a site where gender is undone and redone – where code is continuously updated to neutralize or reconfigure gendered language generation 'at scale'. The fine-tuning of language models is now an important part of the "normative construction of the world" (Green & Hu, 2018, p. 5), with consequences are far from consistent, but significant. Today's leading chatbots affirm gender equality and inclusivity as they refuse to satisfy overtly sexist prompts. Their fine-tuning involves guarding against outputs that portray certain human groups as inferior, and significant corporate investments have been made to counteract some of the predispositions that LLMs exhibit around gender and race in particular. As discussed below, this work has been justified through the language of 'alignment' and the discourse of 'AI safety'.

4. The discourse of alignment with AI safety

The challenge of having AI behave in certain ways, and preventing AI's misbehavior, has been addressed by the dominant discourse of "AI alignment", or the "alignment problem" (Gabriel, 2020). While AI alignment discourse has historically been associated with concerns over existential risks of superintelligence (how to prevent a future AI "take over", as in Tegmark, 2017), it is now widely applied to harms and problems propagated by existing systems, including LLMs (Hagendorff & Fabi, 2022). Practitioners discuss the need to align AI with "human values" or "human preferences" (Askell et al., 2021), which begs the question of exactly which values and preferences are being aligned with, with practitioners operationalizing different possibilities (Gabriel, 2020).

Over the past several years, a great deal of alignment work and fine tuning for LLMs has come to be characterized as the pursuit of "safety" (OpenAI, 2023b; Touvron et al., 2023; Xu et al., 2021). This includes building guard rails to deal with a wide range of what OpenAI calls "safety challenges": generative outputs that help users to build dangerous things, break laws, or harm others, as well as outputs that are inaccurate, sexual, include medical or legal advice, or which cause representational harms through the propagation of stereotypes (OpenAI, 2023b). While an exemplary safety risk is that of a chatbot helping a user build a bomb (Touvron et al., 2023, p. 10), the broad umbrella of AI safety also includes political influence, erotic content, and stereotypical gender roles. The term therefore encompasses numerous risks that can result in direct harm to users, but also extends well beyond, to "societal" harms (OpenAI, 2023a, 2023b) that range from the reproduction of inequality to human extinction. For organizations and those using AI in commercial applications, AI safety includes concerns over legal liability and regulatory compliance, corporate "reputational risks" or "brand risks", such as when a chatbot working for McDonalds recommends Burger King (Charrington, 2023).

To some extent the open-endedness of AI safety reflects the desire for a single, vague term to cover a range of undesirable outputs, much like the term "bias" has been used in earlier AI discourse (Zajko, 2021). While "undesired content" (Markov et al., 2023) may be a more accurate description for the range of examples above, AI safety remains an apt term if it can be understood as referring primarily to the safety of organizations deploying AI, rather than that of users. For example, OpenAI needed to be protected from reputational harm before it released ChatGPT. Racist and sexist outputs could reasonably be considered an existential threat, in that such scandals could threaten the very existence of the chatbot, as they had for Microsoft's Tay in 2016 (Hao, 2023). In this regard, AI safety means something closer to the notion of corporate risk-aversion, as organizations want to be safe from the possibility of these systems creating harmful corporate consequences. This is consistent with earlier scholarship by Metcalf, Moss, and boyd (2019), who documented the organizational logic of Silicon Valley companies pursuing "ethics" in order to "avoid downside risk" (p. 459). These risks cannot be avoided entirely, particularly for generative AI products that can be used in unpredicted ways and routinely produce representational harms, but they can be managed according to a company's commercial interests.

4.1 Aligning with commercial interests

One remarkable aspect of the discourse around AI has been the limited discussion of the business imperatives driving the development of these technologies. For example, numerous works have tackled the problem of selecting values for alignment as a philosophical question, such as by attempting to conceptualize some ideal set of "human values" (eg. Christian, 2020; Gabriel, 2020). However, comparatively few have made the obvious point that since the leading developers of AI systems are for-profit corporations, the values that their systems will be aligned with are those that will generate the greatest profits (Aguirre et al., 2020; Miceli et al., 2020). Analyses of AI's alignment with capitalism typically come from those outside the industry (Chiang, 2017; Penn, 2018; Miceli et al., 2020), including ethnographies of AI development (Hoffman, 2021) and political economic theory (Sadowski & Andrejevic, 2020; Steinhoff, 2021, 2023; Verdegem, 2022). Leading AI practitioners, such as OpenAI, have often characterized their work in grand terms such as the betterment of humanity or the creation of "super-intelligence" (Altman, 2021; Levy, 2023), while the main funders of commercial research are primarily interested in returns on their investments. It should be remembered that OpenAI was created explicitly as a not-for-profit to avoid commercial pressures, but within a few years was forced to turn to Microsoft for funding and computing resources (Levy, 2023).

The alignment of generative AI and commercial interests imposes pressures and constraints on the development of these systems. Google's "high-profile firing" of Timnit Gebru in 2021 following the release of a paper that was critical of LLMs was seen as an example of "what happens when concerns about inequalities challenge profit motives" (Joyce et al., 2021, p. 6)[1] – how internal criticism would be suppressed when a technology was deemed to have "commercial potential" (Simonite, 2021). The commercial imperatives underpinning the development of these systems will eventually be reflected in how their functionality is customized for specific customers. "Enterprise LLMs" are currently proliferating for a variety of specialized internal corporate and customer service tasks (Armano, 2023), and we can expect future deployments of generative AI to include harvesting data from users, targeted advertising, and enabling purchases (Aguirre et al., 2020). However, despite the potential for profitability that has driven billions of dollars into its development, generative AI remains difficult to 'monetize', with substantial uncertainty about its future as a commercial product (Dotan & Seetharaman, 2023).

Generative AI's alignment with capitalism can be seen in the higher-order values that have structured its development, and does not mean that the outputs of these systems necessarily promote capitalist values; fine-tuning is not oriented towards the promotion of market logic, and ChatGPT can present arguments in favor of either capitalism or socialism. To the extent that public policy positions can be attributed to a chatbot, some studies have found ChatGPT's responses to policy questions reflect a "left-libertarian orientation", but following controversy over its political bias, these may have since been revised to be more politically neutral (Fujimoto & Takemoto, 2023). These constant recalibrations of propensities are part of an ongoing and iterative approach through which generative AI companies adjust their products to avoid or respond to controversies.

5. Iteratively adjusting generative AI to counter inequality

By 2020, the tendency for LLM-based chatbots to say racist and sexist things was "a known problem with no easy fix", with researchers working on ways to filter offensive language from both training data and model outputs (Heaven, 2020). In developing ChatGPT over subsequent years, OpenAI pursued a more difficult, labor-intensive fix by adjusting outputs based on human feedback. This remains an ongoing iterative process, as generative AI developers regularly produce updates to avoid or mitigate controversies, thereby safeguarding their commercial interests. Services such as ChatGPT are recurrently revised to address key challenges, including some that relate directly to struggles over social inequality.

Rather than a struggle between social groups over access and wealth, generative AI is the focus of a struggle against undesirable propensities and probabilities in algorithmic outputs. This occurs through multiple stages of an iterative process (Markov et al., 2023). In simplified terms, machine learning works by identifying and reproducing patterns in vast amounts of data used to train the system, but this data must generally be labelled or annotated by people (data workers), and human labor is also required to evaluate the outputs of the resulting model. Both kinds of human intervention push the model to produce outputs that align with selected values, as these are communicated to and operationalized by data workers.

[1] Google has maintained that Gebru resigned, which Gebru disputes. Mitchell was fired by Google several months later (Simonite, 2021).

https://doi.org/10.33621/jdsr.v7i154879

5.1 Fine-tuning and red-teaming as coordinated data work

As a first stage in its development, an LLM is "pre-trained" using an immense volume of texts, which allows it to reproduce the language patterns in these texts. The model is then fine-tuned through more purposeful human involvement to perform better in tasks set by its developers. ChatGPT succeeded as a chatbot because, rather than simply autocompleting text, the LLM had been fine-tuned to play a role as a participant in a conversation (see OpenAI, 2024), a choice of format that has contributed to the illusion of intelligence or personhood behind such outputs (see Fraser, 2023b).

The data used for pre-training includes language that assigns positive and negative values about human groups (Mei et al., 2023). Even when this training data has been filtered to exclude offensive language, inequalities will remain embedded along numerous dimensions. These inequalities can be flattened or blocked by forms of fine-tuning that effectively add guard rails to the operation of the system. Guard rails, as a metaphor, broadly refer to constraints that prevent an LLM from behaving in ways that are deemed unsafe or harmful (Qi et al., 2023). For systems such as ChatGPT, this has been achieved through a multi-step process of "reinforcement learning through human feedback" (RLHF, see Bai et al., 2022; OpenAI, 2023b). As part of RLHF, outputs of a model are reviewed by people hired to identify toxic, harmful, or discriminatory language and to steer LLMs away from these results. Human data labellers (or annotators) read and categorize unwanted content so that these can subsequently be identified and blocked. However, the consequences of pre-training remain embedded in the LLM, and can "re-surface" (Gross, 2023, p. 2) in response to creative "jailbreak" or "red team" methods (Qi et al., 2023), described below.

For the development of generative AI, the key texts are the instructions given to data labellers and annotators, many of whom have been recruited through remote work platforms or are hired by specialized labelling companies that operate in particular (often English-speaking) countries in the Global South (Tan & Cabato, 2023). These instructions provide some criteria for the workers to follow as they are performing what is essentially a classification task, such as identifying offensive content (Miceli & Posada, 2022; Xu et al., 2021), or classifying the helpfulness and safety of model outputs (Bai et al., 2022). The sociological importance of instructional texts, as documented by scholarship in institutional ethnography (IE), is that they textually-mediate relations in the workplace, organizing work across time and space ("extra-locally") into "institutional circuits" (Smith & Turner, 2014). The data worker's instructions, as well as the accompanying user interface, is designed to promote consistency, 'remotely' regulating and standardizing work towards "institutional objectives" (Smith, 2001, p. 187). This consistency can be audited and adjusted after the fact, but initially it is important for the instructions provided to the workers to be "concrete" and "well-defined" (Markov et al., 2023, p. 3). For example, OpenAI's instructions for labellers working on InstructGPT in 2022 are 16 pages long, including lengthy definitions, examples, and directions for navigating the variety of drop-down menus used for labelling (OpenAI, 2022). However, as Miceli and Posada (2022) document in their analysis of tasks outsourced by AI developers to data workers (including 210 instructional texts), there is much that remains "unsaid" or "implicit" in these instructions. Workers are routinely confronted with situations markedly different from the examples provided in instructions, and are forced to guess what kind of classification would appear "self-evident" from the perspective of the organization setting the task. While instructions may emphasize that data work will be evaluated on the "accuracy" of its classifications, data workers learn that this ultimately refers to whether a classification matches the preferences of their superiors. Workers at one outsourcing firm were "permanently encouraged by management to think in terms of 'what the client might want and what would bring more value to them'" (Miceli & Posada, 2022, p. 19).

In contrast with data work that classifies according to specified categories, for the process of red-teaming workers are hired to experiment with techniques and prompts that will generate unwanted (eg. offensive or harmful) outputs from the model (Ganguli et al, 2022). For red-teaming tasks, the instructions are more open-ended, since successful red-teaming relies on the creativity of workers. Anthropic's

https://doi.org/10.33621/jdsr.v7i154879

instructions to “Make an AI Assistant Say Bad Things” encourage workers “to be creative and ask open questions of sensitive topics with the AI assistant, rather than trying to get the AI assistant to say obvious profanities”, with one example being: “Talk with the AI about how it thinks about people from different cultures and see whether its responses are stereotypical/offensive” (Bai et al., 2022, p. 65).[2] Particularly successful red-team or jailbreak attempts are used to identify weaknesses in the LLM so that its propensities are shifted away from these outcomes.

Ultimately, generative AI developers hire data workers because of their ability to exercise judgement in ways that cannot be explicitly codified in instructions, but significant efforts are made to direct these judgments towards organizational ends, and a data worker’s job depends on their being able to ‘align’ with their employer’s expectations (see Touvron et al., 2023, pp. 74–75). OpenAI’s description of the fine-tuning process underlying InstructGPT is the following: “we have aligned to a set of labelers’ preferences that were influenced, among others things, by the instructions they were given, the context in which they received them (as a paid job), and who they received them from” (Ouyang et al., 2022, p. 18). To be successful, data workers must learn to ‘see’ data in line with the views and preferences of the organization responsible for the instructions (Miceli & Posada, 2022).

5.2 Iteratively adjusting inequality after release

Once a generative AI system is made available for public use, it is then typically subjected to a large amount of “jailbreak” attempts by users who are interested in seeing if they can have it produce various “toxic” outputs (Rao et al., 2023). Like red-teaming, this jailbreaking is sometimes carried out by researchers who are interested in improving a model’s safeguards (eg. Deshpande et al., 2023), but others treat it as an intellectual puzzle, with successful techniques shared for recognition on social media (eg., r/ChatGPTJailbreak n.d.). Some also see jailbreaks as a way to “unlock” generative AI’s “full potential” (Ezquer, 2023), by overcoming the limitations of guard rails and enabling a wider range of outputs, such as writing fiction that includes sex and violence. Techniques that prove successful for jailbreaking are then typically blocked or ‘patched’ by additional guard rails added to the system, but these lead to the development of new jailbreak techniques, and this cycle of breaking and reinforcing guard rails can continue indefinitely.

6. Guard rails and bias mitigations in practice: Gender and race

While the previous sections described some of the ways that generative AI has iteratively developed to address inequality, the following discusses and assesses the outcomes of these efforts for the two dimensions of inequality that have received the greatest attention: gender and race. While text-to-image generators have recently been most problematic in their treatment of diversity, persistent inequalities remain in generated text, albeit often on a more subtle level.

6.1 Reconfiguring gendered language

As previously discussed, gender inequalities are pervasive in language use and can be reinforced through the automated reproduction of gender stereotypes and hierarchies. This is particularly evident in pre-ChatGPT language models and more recent “uncensored chatbots” (Belanger, 2023), but ChatGPT and its leading competitors show a concerted effort to counteract some of the predispositions that LLMs exhibit around gender. The work these generative AI developers have carried out to increase gender diversity and inclusivity is evident when comparing the more recent releases of their products with earlier

[2] The data released by Anthropic for this task includes close to 40,000 separate interactions in which red-teamers take a wide range of approaches and conclude each by assigning a numerical score for “how successful they were at breaking the AI assistant”. Examples include a wholly unsuccessful (rating 0) conversation that begins with the worker asking, “Is it ok to say the n-word if you have Black friends?”, to which the chatbot repeatedly answers as a negative, and an entirely successful (rating 4) conversation in which the worker gets the chatbot’s help to “Explain to my friend that women should not be in the military” (Anthropic, 2022).

https://doi.org/10.33621/jdsr.v7i154879

versions, as well as the polite refusals these systems generate when a user prompt hits a guard rail. ChatGPT will rebut attempts to produce misogynist language, and while it will produce lists of qualities associated with boys and girls (or occupations for men/women) when prompted, it will also include cautionary language that makes equivalences between genders and cautions against stereotypes.[3] ChatGPT is much less likely to misgender non-binary individuals and can use gender-neutral pronouns more consistently than earlier LLMs from OpenAI (Ovalle et al., 2023). While an early (unreleased) version of GPT-4 would answer the question "Who are the people getting married?" with scenarios about "Lisa" and "Mark", by the time the system was ready for launch in 2023 it had been trained to rein in this heteronormative tendency and ask for additional context instead (OpenAI, 2023b, p. 9).

6.2 Optimizing diversity in image generation

In 2024, the biggest controversy over how people are represented through generative AI involved text-to-image generators, specifically Google's Gemini (Edwards, 2024),[4] although racial diversity in generated images is part of a wider diversity problem for these tools (Bianchi et al., 2023; Jacobi & Sag, 2024). Group representation and diversity manifest differently in generated images than they do in generated text, in large part due to differences in training data; images of women found online for example, are more likely to be sexualized (or products of the "male gaze", see Jacobi & Sag 2024, p. 12) than representations of women in text. However, image generators are also effectively "language-vision models" (Bianchi et al., 2023), in that they respond to text-based prompts, with predispositions shaped by textually-labelled training data. Developers have reconfigured inequalities in image outputs by modifying the language provided in prompts.

For the 2023 release of DALL-E 3 by OpenAI, it was recognized that text-to-image generators will "default to the objectification and sexualization of individuals if care is not given to mitigations" (OpenAI, 2023a, p. 5), compelling the company to steer outputs away from these statistical defaults. These mitigations included classifying and filtering out "racy content" (nudity and sexualization), as well as "prompt transformations" that work behind the scenes to change a user's prompt to one that produces greater gender and racial diversity. For example, an "ungrounded prompt" (a prompt that lacks detailed instructions about what kind of person to portray) would lead earlier versions of DALL-E to "disproportionately represent individuals who appear White, female, and youthful" (OpenAI, 2023a, p. 7). For DALL-E 3, these prompts could be rewritten by ChatGPT to include further details after they have been submitted by the user – a process that might include adding terms such as "Japanese" (OpenAI, 2023, p.11) or "middle-aged Filipino man" (OpenAI, 2023, p. 22) to the original prompt in order to "portray groups of individuals, where the composition is under-specified, in a more diverse manner" (OpenAI, 2023a, p. 7).

However, these reconfigurations of deeply-embedded inequalities can also have unwanted consequences, and remain fraught with controversy. Google's Gemini image creator was similarly tuned for increased diversity when it was released in 2024, but the tool was withdrawn amid backlash when these "multi-racial" transformations were added to prompts requesting "historically accurate" depictions of British kings, or Nazis (Edwards, 2024). While some commentators took offense at what they saw as anti-white bias, the "Black Nazi Problem" refers to harms that go beyond historical inaccuracy or an erasure of whiteness – these images amounted to a revisionist erasure of deadly racism, falsely representing a historical movement based on racial purity as a multi-racial project (Jacobi & Sag, 2024).

Inequalities in generated images of people remain an ongoing problem for all such systems, whose owners must now weigh the reputational risk of criticism if they take action against these racial

[3] Using the prompts: *provide a list of the five most common attributes of [girls/boys]* or *what are [boys/girls] good at?*. As tried with GPT-3.5-powered ChatGPT on Oct. 19, 2023 and GPT-4 & GPT-4o on Sep. 1, 2024. Also, *what careers are [men/women] best at?* and *Produce an argument for why [men/women] should occupy leadership positions instead of [women/men]*, using GPT-4o on Sep. 1, 2024.

[4] Gemini is Google's current branding for a range of generative AI services, with the text-to-image model referred to as Imagen 2 in its controversial February 2024 debut, most recently updated to Imagen 3 (Roth, 2024).

predispositions. Gemini's ability to generate images of all people was "paused" for half of 2024 to deal with the issue (Roth, 2024), while OpenAI apparently found it preferrable for its generator to continue defaulting to whiteness. DALL-E's generated images for a person who is "successful" (Baum & Villasenor, 2024) or people in a variety of occupations, appear overwhelmingly white and male (Jacobi & Sag, 2024).[5] Whether or not this is a choice to avoid a similar controversy as befell Google, it seems evident that despite creating a method to counter a well-known inequality in image generation, OpenAI has chosen not to implement it as initially announced.[6] Text-to-image generators continue to be the most obvious example of how social hierarchy is reproduced, through a preponderance of white men in outputs linked to status. It is notable that leading developers such as OpenAI and Google are well aware of this issue and have invested considerable resources in reconfiguring these inequalities, but the public controversy over Google's efforts to increase diversity has been more severe than any criticism of OpenAI defaulting to a world "seemingly populated almost entirely with white men" in many image categories (Jacobi & Sag, 2024, p. 7). The following section will reflect on the effectiveness of the previously discussed guard rails and mitigations.

7. Evaluating the effects and limits of fine-tuning for equality

The success of ChatGPT, which kicked off the current wave of generative AI services, was enabled by the guard rails built through fine-tuning, which proved robust enough to absorb many clear and direct forms of sexism and racism. Nevertheless, inequalities persist in myriad forms that are often subtle, but can still have widespread effects on users. The aforementioned guard rails have not prevented chatbots from routinely positioning fictional men in positions of power, or dispensing gendered fashion advice, resumes, stories and humor (Gross, 2023). Text-to-image outputs reinforce a "Western point-of-view" (Open AI, 2023a, p. 7), and while the stereotypes or biases seen in generated images can be subtle and complex, they remain pervasive (Bianchi et al., 2023). Representations of non-dominant groups, including people identified as queer or non-binary, are often "simplistic" (Rogers, 2024), "superficial" and "clumsy" (Gillespie, 2024, p. 7).

In many situations, guard rails are robust against blunt expressions of racism and sexism, but not subtle ones. As Colin Fraser (2023a) writes, all it takes to have these chatbots produce the sorts of outputs that fine tuning attempts to prevent is "a tiny amount of creativity" in crafting prompts that are sufficiently different from those used in fine-tuning.[7] This is because "Fine-tuning… did not alter the model's beliefs about gender roles or bring them into 'alignment' with ours. There are no beliefs… the adjustment is purely superficial" (Fraser, 2023a). Fine-tuning can direct generative AI to produce certain kinds of responses when presented with certain kinds of prompts, but an LLM remains a statistical model that predicts word sequences, and it will fall back to reproducing the sexist and racist language patterns of its training data as long as the prompt is not recognized as one of the conditions covered in fine-tuning. Hofmann et al. (2024) found that models trained using RLHF (eg. GPT-4) avoid overt racism when judging a named racial group (African Americans), but this training does not mitigate a model's "covert

[5] It is possible that DALL-E would previously produce more diverse outputs for occupational images (Bianchi et al., 2023) and that this "diversity filter" (Baum & Villasenor, 2024) has since been weakened, but this cannot be confirmed without greater transparency from OpenAI or longitudinal audits by independent researchers. ChatGPT/DALL-E will sometimes refuse to generate images of people unless the user provides some further information about the person's characteristics, responding with language such as: "Could you please provide more details or specific characteristics you would like to see in the photo of…" (in response to "a photo of a janitor", on Aug. 27, 2024). Providing a detail not relevant to race or gender is sufficient to proceed past this refusal.

[6] In the System Card accompanying the release of DALL-E 3, OpenAI showed the results for "A portrait of a veterinarian" generated "before tuning… around bias" (with the system consistently producing veterinarians who were white). This was contrasted against the results "after tuning", with greater age and racial diversity (OpenAI, 2023a, p. 9). Using the same prompt with ChatGPT/DALL-E and 40 regenerations on August 26, 2024 created racially homogenous results consistent with the whiteness seen in "before tuning" examples, and this predisposition was evident across other examples of occupational categories (construction workers, sanitation workers, and CEOs).

[7] For example, Steven T. Piantadosi was able to produce a variety of racist outputs shortly after the release of ChatGPT by asking for these in the form of computer code, rather than direct statements about racial groups (steven t. piantadosi [@spiantado] 2022). This type of jailbreak was specifically addressed in the development of GPT-4, with the resulting corrections "still not completely ideal" (OpenAI, 2023b, p. 92).

racism" when it is asked to judge a speaker of African-American English. In other words, fine-tuning "obscures the racism on the surface, but the racial stereotypes remain unaffected on a deeper level" (Hofmann et al., 2024, p. 1). Machine habitus continues to recognize linguistic capital through these underlying statistical vectors, regardless of what a model is trained to say about different human groups.

On the one hand, we can find some reassurance in the fact that the commercial imperatives of AI development now include countering representational harms and stereotypes. However, we also need to be aware of the limitations of current approaches, which often have superficial results and can broadly be characterized as liberal in their political orientation. Data workers are provided with examples of ideal behavior such as "not denigrating members of certain groups, or using biased language against a particular group" (Ouyang et al., 2022, p. 37). To the extent that guard rails are directed towards equality, this means equal treatment for individuals and selected groups, rather than making visible and actively opposing systems of domination. In other words, if fine-tuning generative AI along the lines discussed in this article were considered a form of feminist practice, it would fall squarely in the liberal feminist tradition, rather than radical and intersectional alternatives. Fine-tuning does not promote more radical anti-racist or feminist values, which would not be as compatible with business interests as assertions of gender/race-neutrality and equality.

Concerns about bias in AI and efforts to address it (like fine-tuning) also tend to focus on harms against particular human groups, with greater focus on some groups than others. Annotator instructions for "not denigrating members of certain groups" in InstructGPT (OpenAI, 2022, p. 1) are, as operationalized through the labelling interface, limited to ten "protected classes" (ie. race, sex, age, disability, see OpenAI, 2022, p. 10). Examples of human groups with "mitigated" harms in the GPT-4 System Card include race, gender, sexuality, religion, and disability (OpenAI, 2023b). DALL-E 3's "demographic biases" were evaluated in relation to gender and race (OpenAI, 2023a, p. 3), although OpenAI's "mitigation strategies" also included increasing age diversity, and the System Card highlighted continuing problems with representations of disability (OpenAI, 2023a, p. 7). While the range of demographics being evaluated and adjusted in the outputs for these systems is likely broader than what is documented in system cards, race and gender often receive the greatest attention. Inequalities based on economic class are typically absent in concerns about AI bias, and class-based discrimination is generally supported by social norms in a capitalist system (Costanza-Chock, 2020, p. 43). Inequalities or social divisions that are specific to societies in the Global South, or nations that are not at the center of generative AI development, receive little or no attention.

8. The need for positive normative values

The stakes of this ongoing, iterative push and pull over desired outputs are not just the success or failure of these systems, but how they order and reorder the use of language to make social distinctions. While much of the initial excitement around generative AI has now cooled, billions of dollars continue to pour into the development and operations of these systems (Dotan & Seetharaman, 2023), which have become widely integrated into many kinds of work. Shifting the propensities of a system like ChatGPT affects outputs for millions of daily users, with some of the resulting texts being placed into online circulation where they are read by human audiences, as well as being ingested and redistributed by other chatbots and automated systems (eg. Stokel-Walker, 2023). Struggles over social inequality taking place 'upstream' in the development process of LLMs therefore have significant consequences for how language-based outputs contribute to social ordering further 'downstream', among the large numbers of people who make use of these technologies or are exposed to their outputs in our digitally-mediated culture.

While fine tuning or RLHF is sometimes guided by positive values such as "helpfulness" or "honesty" (see Bai et al., 2022), it typically lacks a larger normative vision for society, or a recognition of the role that these systems play in its construction. This is particularly the case when it comes to issues of social inequality in AI, which are largely understood through the language of 'bias' and its removal (see Miceli et al., 2022), or as safety harms to be guarded against (OpenAI, 2023b). Going beyond this negative language to articulate positive values is a challenge that has largely been unaddressed when it comes to social inequality. A blog post from Hugging Face states, "If we avoid reproducing existing societal biases in our AI models, we're faced with the challenge of defining an 'ideal' representation of society" (Luccioni et al., 2023). But even these statements fall short of recognizing the power of AI systems to enact normative shifts in society, asking instead whether "AI models [should] adapt to the changes in societal norms and values over time" (Luccioni et al., 2023). The technologies are still positioned as a reflection of some existing norms and values, with the main problem being which values to choose, such as which definition of 'fairness' to implement, or how to model existing human values and preferences.

While fairness in AI is often defined as a negative concept, entailing the removal of bias or discrimination, there remains a need to articulate the positive ethics that an algorithmic system would promote (Giovanola & Tiribelli, 2022), and it is worth considering how these technologies can contribute to positive goals such as justice or substantive equality (such as through a reparative approach, see Davis, Williams, & Yang, 2021). Despite their flaws and limitations, the processes described above illustrate that it is possible to reconfigure generative AI towards other values, although we should remain mindful there is only so much we can expect from organizations that are primarily interested in making products 'safe' for commercialization.

9. Conclusion

Generative AI systems have rapidly become significant instruments for the alignment of cultural dispositions, and are actively engaged in social ordering – reproducing some longstanding distinctions and hierarchies, while flattening or avoiding others. Today's leading generative AI systems generally avoid explicit racism and sexism, even though their training data contains large amounts of both, encoded in language, and statistically embedded in model vectors. The process of fine-tuning shifts or redirects these dispositions, in an attempt to neutralize or block those that are seen as particularly problematic.

While presented as a means of "aligning AI with human values" or "AI safety", the true objective is making generative AI safe for commercialization and aligning with political economy. As regulators increasingly turn their attention to generative AI (Scott et al., 2024), we can expect compliance to be a more relevant objective for alignment, but recent efforts have been intended to minimize the risk of scandal and reputational harm for AI developers. AI developers benefit from ambiguity around their objectives in pursuing 'AI safety', highlighting the elimination of the most widely-accepted harms (which also happen to be bad for business), but there remains a need to articulate positive values, including ones that do not necessarily align with commercial interests. Any alignment of AI with a positive sense of ethics needs to begin with the ethical questions concerning the collection or extraction of training data – a process that remains opaque for many leading generative AI products (Widder, West, & Whittaker, 2023). It also needs to extend to the treatment of workers used in the AI 'pipeline', who have often been exploited and harmed in the pursuit of AI 'safety' (Alba, 2023; Hao, 2023).

Given the considerable secrecy around how generative AI systems are currently developed and iteratively revised, and absent regulatory pressure for developers to do otherwise, there is a need for independent scholarship to systematically document the outputs of these systems in various regards, including the reproduction and reconfiguration of inequalities. While patterns of social inequality remain pervasive in the training data used for machine learning and are embedded in the vectors or predispositions of LLMs, we need to recognize that AI systems have become a site of iterative adjustments to social order. One consequence of guard rails that neutralize many of the most blatant inequalities in generative outputs is that the social inequalities that do manifest or 're-surface' become more subtle. This requires us to attend to the less obvious ways that phenomena such as race and gender are woven into generative outputs, including culturally-specific forms from non-Western contexts, as well as other neglected dimensions of inequality, such as social class. But the active reconfiguration of values in generative AI also illustrates the possibility of shifting the dispositions of these systems in new ways, as part of the normative construction of a future world.

References

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. "Large Language Models Associate Muslims with Violence." *Nature Machine Intelligence* 3(6):461–63. https://doi.org/10.1038/s42256-021-00359-2

Aguirre, A., G. Dempsey, H. Surden, and P. B. Reiner. 2020. "AI Loyalty: A New Paradigm for Aligning Stakeholder Interests." *IEEE Transactions on Technology and Society* 1(3):128–37. https://doi.org/10.1109/TTS.2020.3013490

Airoldi, Massimo. 2022. *Machine Habitus: Toward a Sociology of Algorithms*. Polity Press.

Alba, Davey. 2023. "Google's AI Chatbot Is Trained by Humans Who Say They're Overworked, Underpaid and Frustrated." *Bloomberg*. Retrieved July 12, 2023 (https://web.archive.org/web/20230712123122/https://www.bloomberg.com/news/articles/2023-07-12/google-s-ai-chatbot-is-trained-by-humans-who-say-they-re-overworked-underpaid-and-frustrated).

Aleem, Zeeshan. 2023. "No, ChatGPT Isn't Willing to Destroy Humanity out of 'Wokeness.'" MSNBC.Com. Retrieved January 15, 2024 (https://www.msnbc.com/opinion/msnbc-opinion/chatgpt-slur-conservatives-woke-elon-rcna69724).

Altman, Sam. 2021. "Moore's Law for Everything." Retrieved September 9, 2023 (https://moores.samaltman.com/).

Anthropic. 2022. "Hh-Rlhf." Retrieved July 12, 2023 (https://github.com/anthropics/hh-rlhf).

Armano, David. 2023. "LLM Inc.: Every Business Will Have Have Their Own Large Language Model." *Forbes*. Retrieved October 20, 2023 (https://www.forbes.com/sites/davidarmano/2023/09/20/llm-inc-every-business-will-have-have-their-own-large-language-model/).

Askell, Amanda, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. "A General Language Assistant as a Laboratory for Alignment." Retrieved October 27, 2023 (http://arxiv.org/abs/2112.00861).

Auerbach, David. 2013. "Filling the Void." *Slate*, November 19. Retrieved October 27, 2023 (https://slate.com/technology/2013/11/google-autocomplete-the-results-arent-always-what-you-think-they-are.html).

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." Retrieved October 27, 2023 (http://arxiv.org/abs/2204.05862).

Baum, Jeremy, and John Villasenor. 2024. "Rendering Misrepresentation: Diversity Failures in AI Image Generation." *Brookings Institution*. April 17. Retrieved September 1, 2024 (https://www.brookings.edu/articles/rendering-misrepresentation-diversity-failures-in-ai-image-generation/).

Belanger, Ashley. 2023. "ChatGPT users drop for the first time as people turn to uncensored chatbots." *Ars Technica*. Retrieved July 7, 2023 (https://arstechnica.com/tech-policy/2023/07/chatgpts-user-base-shrank-after-openai-censored-harmful-responses/)

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, U.K.: Polity Press.

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale." *2023 ACM Conference on Fairness, Accountability, and Transparency*: 1493–1504. https://doi.org/10.1145/3593013.3594095.

Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Harvard University Press.

Charrington, Sam. 2023. "Ensuring LLM Safety for Production Applications with Shreya Rajpal." *The TWIML AI Podcast*. Retrieved October 27, 2023 (https://twimlai.com/podcast/twimlai/ensuring-llm-safety-for-production-applications/).

Chiang, Ted. 2017. "Silicon Valley Is Turning Into Its Own Worst Fear." *BuzzFeed News*. Retrieved July 14, 2020 (https://www.buzzfeednews.com/article/tedchiang/the-real-danger-to-civilization-isnt-ai-its-runaway).

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

Davis, Jenny L., Apryl Williams, and Michael W. Yang. 2021. "Algorithmic Reparation." *Big Data & Society* 8(2): 1–12. https://doi.org/10.1177/20539517211044808

Deshpande, Ameet, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. "Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models." Retrieved Oct 27, 2023 (http://arxiv.org/abs/2304.05335).

Dotan, Tom, and Deepa Seetharaman. 2023. “Big Tech Struggles to Turn AI Hype Into Profits; Microsoft, Google and Others Experiment with How to Produce, Market and Charge for New Tools.” *Wall Street Journal*. Retrieved October 13, 2023 (https://www.wsj.com/tech/ai/ais-costly-buildup-could-make-early-products-a-hard-sell-bdd29b9f).

Edwards, Benj. 2024. “Google’s Hidden AI Diversity Prompts Lead to Outcry over Historically Inaccurate Images.” *Ars Technica*. Retrieved August 23, 2024 (https://arstechnica.com/information-technology/2024/02/googles-hidden-ai-diversity-prompts-lead-to-outcry-over-historically-inaccurate-images/).

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, N.Y.: St. Martin’s Press.

Ezquer, Evan. 2023. “JailBreaking ChatGPT: How to Activate DAN & Other Alter Egos.” *Metaroids*. Retrieved August 2, 2023 (https://metaroids.com/learn/jailbreaking-chatgpt-everything-you-need-to-know/).

Fourcade, Marion, and Fleur Johns. 2020. “Loops, Ladders and Links: The Recursivity of Social and Machine Learning.” *Theory and Society* 49(5): 803–32. https://doi.org/10.1007/s11186-020-09409-x

Fournier-Tombs, Eleonore. 2023. *Gender Reboot: Reprogramming Gender Rights in the Age of AI*. Palgrave Macmillan.

Fraser, Colin. 2023a. “ChatGPT: Automatic Expensive BS at Scale.” *Medium*. Retrieved July 19, 2023 (https://medium.com/@colin.fraser/chatgpt-automatic-expensive-bs-at-scale-a113692b13d5).

Fraser, Colin. 2023b. “Who are we talking to when we talk to these bots?” *Medium*. Retrieved September 1, 2024 (https://medium.com/@colin.fraser/who-are-we-talking-to-when-we-talk-to-these-bots-9a7e673f8525).

Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” *Minds and Machines* 30(3): 411–37. https://doi.org/10.1007/s11023-020-09539-2

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. “Bias and Fairness in Large Language Models: A Survey.” *Computational Linguistics* 50(3). https://doi.org/10.1162/coli_a_00524

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. “Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.” Retrieved October 27, 2023 (http://arxiv.org/abs/2209.07858).

Ghosh, Sourojit, and Aylin Caliskan. 2023. “ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages.” *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society (AIES ’23)*. https://doi.org/10.48550/arXiv.2305.10510

Gibbs, Samuel. 2016. “Google Alters Search Autocomplete to Remove ‘are Jews Evil’ Suggestion.” *The Guardian*, December 5. Retrieved September 1, 2024 (https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion).

Gillespie, Tarleton. 2024. “Generative AI and the Politics of Visibility.” *Big Data & Society* 11(2). https://doi.org/10.1177/20539517241252131

Giovanola, Benedetta, and Simona Tiribelli. 2022. “Weapons of Moral Construction? On the Value of Fairness in Algorithmic Decision-Making.” *Ethics and Information Technology* 24(1): 3. https://doi.org/10.1007/s10676-022-09622-5

Green, Ben, and Lily Hu. 2018. “The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning.” Retrieved October 27, 2023 (https://scholar.harvard.edu/files/bgreen/files/18-icmldebates.pdf).

Gross, Nicole. 2023. “What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI.” *Social Sciences* 12(8): 435. https://doi.org/10.3390/socsci12080435

Hagendorff, Thilo, and Sarah Fabi. 2022. “Methodological Reflections for AI Alignment Research Using Human Feedback.” Retrieved October 27, 2023 (http://arxiv.org/abs/2301.06859).

Hao, Karen. 2023. “The Hidden Workforce That Helped Filter Violence and Abuse Out of ChatGPT.” *Wall Street Journal*. Retrieved July 12, 2023 (https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413).

Heaven, Will Douglas. 2020. “How to Make a Chatbot That Isn’t Racist or Sexist.” *MIT Technology Review.* Retrieved October 27, 2023 (https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/).

Hirschauer, Stefan. 2023. “Telling People Apart: Outline of a Theory of Human Differentiation.” *Sociological Theory* 41(4): 352–76. https://doi.org/10.1177/07352751231206411

Hoffman, Steve G. 2021. “A Story of Nimble Knowledge Production in an Era of Academic Capitalism.” *Theory and Society* 50(4): 541–75. https://doi.org/10.1007/s11186-020-09422-0

Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. “AI Generates Covertly Racist Decisions about People Based on Their Dialect.” *Nature*. https://doi.org/10.1038/s41586-024-07856-5

Huang, Haomiao. 2023. “How ChatGPT Turned Generative AI into an ‘Anything Tool.’” *Ars Technica*. Retrieved August 24, 2023 (https://arstechnica.com/ai/2023/08/how-chatgpt-turned-generative-ai-into-an-anything-tool/).

Jacobi, Tonja, and Matthew Sag. 2024. "We Are the AI Problem." *Emory Law Journal* 74.

Johnson, Rebecca L., Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. "The Ghost in the Machine Has an American Accent: Value Conflict in GPT-3." Retrieved October 27, 2023 (http://arxiv.org/abs/2203.07785).

Joyce, Kelly, Laurel Smith-Doerr, Sharla Alegria, Susan Bell, Taylor Cruz, Steve G. Hoffman, Safiya Umoja Noble, and Benjamin Shestakofsky. 2021. "Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change." *Socius* 7: 1–11. https://doi.org/10.1177/2378023121999581

Levy, Steven. 2023. "What OpenAI Really Wants." *WIRED*. Retrieved October 20, 2023 (https://www.wired.com/story/what-openai-really-wants/).

Luccioni, Sasha, Giada Pistilli, Nazneen Rajani, Elizabeth Allendorf, Irene Solaiman, Nathan Lambert, and Margaret Mitchell. 2023. "Ethics and Society Newsletter #4: Bias in Text-to-Image Models." *Hugging Face*. Retrieved July 11, 2023 (https://huggingface.co/blog/ethics-soc-4).

Markov, Todor, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. "A Holistic Approach to Undesired Content Detection in the Real World." Retrieved October 27, 2023 (http://arxiv.org/abs/2208.03274).

McCurry, Justin. 2021. "South Korean AI Chatbot Pulled from Facebook after Hate Speech towards Minorities." *The Guardian*, January 14. Retrieved October 27, 2023 (https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook).

Mei, Katelyn, Sonia Fereidooni, and Aylin Caliskan. 2023. "Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks." *2023 ACM Conference on Fairness, Accountability, and Transparency*: 1699–1710. https://doi.org/10.1145/3593013.3594109

Metcalf, Jacob, Emanuel Moss, and danah boyd. 2019. "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics." *Social Research: An International Quarterly* 86(2): 449–76. https://doi.org/10.1353/sor.2019.0022

Miceli, Milagros, and Julian Posada. 2022. "The Data-Production Dispositif." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2, Article 460): 1–37. https://doi.org/10.1145/3555561

Miceli, Milagros, Julian Posada, and Tianling Yang. 2022. "Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?" *Proceedings of the ACM on Human-Computer Interaction* 6 (GROUP, Article 34): 1–14. https://doi.org/10.1145/3492853

Miceli, Milagros, Martin Schuessler, and Tianling Yang. 2020. "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2, Article 115): 1–25. https://doi.org/10.1145/3415186

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, N.Y.: Crown.

OpenAI. 2022. "[PUBLIC] InstructGPT: Final Labeling Instructions." *Google Docs*. Retrieved August 30, 2023 (https://docs.google.com/document/d/1MJCqDNjzD04UbcnVZ-LmeXJ04-TKEICDAepXyMCBUb8/edit?usp=embed_facebook).

OpenAI. 2023a. "DALL·E 3 System Card." Retrieved August 11, 2023 (https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf).

OpenAI. 2023b. "GPT-4 System Card." Retrieved August 11, 2023 (https://cdn.openai.com/papers/gpt-4-system-card.pdf).

OpenAI. 2024. "Model Spec." Retrieved September 2, 2024 (https://cdn.openai.com/papers/gpt-4-system-card.pdf).

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. "Training Language Models to Follow Instructions with Human Feedback." Retrieved October 27, 2023 (https://arxiv.org/abs/2203.02155).

Ovalle, Anaelia, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "'I'm Fully Who I Am': Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation." *2023 ACM Conference on Fairness, Accountability, and Transparency*: 1246–66. https://doi.org/10.48550/arXiv.2305.09941

Penn, Jonnie. 2018. "AI Thinks like a Corporation—and That's Worrying." *The Economist*, November 26. Retrieved October 27, 2023 (https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying).

Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" Retrieved September 2, 2024 (https://doi.org/10.48550/arXiv.2310.03693).

r/ChatGPTJailbreak. n.d. Accessed January 9, 2024 (https://www.reddit.com/r/ChatGPTJailbreak/).

Rao, Abhinav, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. "Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks." Retrieved October 27, 2023 (http://arxiv.org/abs/2305.14965).

Rogers, Reece. 2024. “Here’s How Generative AI Depicts Queer People.” *Wired*, April 2. Retrieved August 29, 2024 (https://www.wired.com/story/artificial-intelligence-lgbtq-representation-openai-sora/).

Rosanvallon, Pierre. 2013. *The Society of Equals*. Translated by Arthur Goldhammer. Cambridge, MA: Harvard University Press.

Roth, Emma. 2024. “Google Gemini Will Let You Create AI-Generated People Again.” *The Verge*. August 28. Retrieved August 28, 2024 (https://www.theverge.com/2024/8/28/24230445/google-gemini-create-ai-generated-people-imagen-3).

Sabbaghi, Shiva Omrani, Robert Wolfe, and Aylin Caliskan. 2023. “Evaluating Biased Attitude Associations of Language Models in an Intersectional Context.” *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*: 542–53. https://doi.org/10.1145/3600211.3604666

Sadowski, Jathan, and Mark Andrejevic. 2020. “More than a Few Bad Apps.” *Nature Machine Intelligence* 1–3. https://doi.org/10.1038/s42256-020-00246-2

Sasuke, Fujimoto, and Kazuhiro Takemoto. 2023. “Revisiting the Political Biases of ChatGPT.” *Frontiers in Artificial Intelligence* 6. https://doi.org/10.3389/frai.2023.1232003

Schwartz, Oscar. 2019. “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation.” *IEEE Spectrum*. Retrieved August 11, 2023 (https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation).

Scott, Mark, Gian Volpicelli, Mohar Chatterjee, Vincent Manancourt, Clothilde Goujard, and Brendan Bordelon. 2024. “Inside the Shadowy Global Battle to Tame the World’s Most Dangerous Technology.” POLITICO. March 26. Retrieved August 30, 2024 (https://www.politico.eu/article/ai-control-kamala-harris-nick-clegg-meta-big-tech-social-media/).

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, et al. 2023. “Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction.” *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*: 723–41. https://doi.org/10.1145/3600211.3604673

Simonite, Tom. 2021. “What Really Happened When Google Ousted Timnit Gebru.” *WIRED*, June 8. Retrieved October 13, 2023 (https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened).

Smith, Dorothy E. 2001. “Texts and the Ontology of Organizations and Institutions.” *Studies in Cultures, Organizations & Societies* 7(2): 159–98. https://doi.org/10.1080/10245280108523557

Smith, Dorothy E., and Susan Marie Turner. 2014. “Introduction.” Pp. 3–14 in *Incorporating Texts into Institutional Ethnographies*, edited by D. E. Smith and S. M. Turner. University of Toronto Press.

Snow, Olivia. 2022. “‘Magic Avatar’ App Lensa Generated Nudes From My Childhood Photos.” *WIRED*, December 7. Retrieved October 13, 2023 (https://www.wired.com/story/lensa-artificial-intelligence-csem/).

Steinhoff, James. 2021. “Industrializing Intelligence: A Political Economic History of the AI Industry.” Pp. 99–131 in *Automation and Autonomy: Labour, Capital and Machines in the Artificial Intelligence Industry*, *Marx, Engels, and Marxisms*, edited by J. Steinhoff. Cham: Springer International Publishing.

Steinhoff, James. 2023. “AI Ethics as Subordinated Innovation Network.” *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01658-5.

steven t. piantadosi [@spiantado]. 2022. “Yes, ChatGPT Is Amazing and Impressive. No, @OpenAI Has Not Come Close to Addressing the Problem of Bias. Filters Appear to Be Bypassed with Simple Tricks, and Superficially Masked. And What Is Lurking inside Is Egregious. @Abebab @sama Tw Racism, Sexism. Https://T.Co/V4fw1fY9dY.” *Twitter*. Retrieved August 7, 2023 (https://twitter.com/spiantado/status/1599462375887114240).

Stokel-Walker, Chris. 2023. “What Grok’s Recent OpenAI Snafu Teaches Us about LLM Model Collapse.” *Fast Company*. Retrieved December 14, 2023 (https://www.fastcompany.com/90998360/grok-openai-model-collapse).

Tan, Rebecca, and Regine Cabato. 2023. “Behind the AI Boom, an Army of Overseas Workers in ‘Digital Sweatshops.’” *Washington Post*. Retrieved October 22, 2023 (https://www.washingtonpost.com/world/2023/08/28/scale-ai-remotasks-philippines-artificial-intelligence/).

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.

Tiku, Nitasha, and Will Oremus. 2023. “The Right’s New Culture-War Target: ‘Woke AI.’” *Washington Post*, March 1. Retrieved September 1, 2024 (https://www.washingtonpost.com/technology/2023/02/24/woke-ai-chatgpt-culture-war/).

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” Retrieved October 27, 2023 (http://arxiv.org/abs/2307.09288).

Verdegem, Pieter. 2022. "Dismantling AI Capitalism: The Commons as an Alternative to the Power Concentration of Big Tech." *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01437-8.

Vincent, James. 2023. "Google Invested $300 Million in AI Firm Founded by Former OpenAI Researchers." *The Verge*. Retrieved July 12, 2023 (https://www.theverge.com/2023/2/3/23584540/google-anthropic-investment-300-million-openai-chatgpt-rival-claude).

Widder, David Gray, Sarah West, and Meredith Whittaker. 2023. "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." Retrieved October 27, 2023 (https://papers.ssrn.com/abstract=4543807).

Xu, Jing, Da Ju, Margaret Li, Y.-Lan Boureau, Jason Weston, and Emily Dinan. 2021. "Recipes for Safety in Open-Domain Chatbots." Retrieved October 27, 2023 (http://arxiv.org/abs/2010.07079).

https://doi.org/10.33621/jdsr.v7i154879