Ethics and Information Technology (2024) 26:42
https://doi.org/10.1007/s10676-024-09783-5

ORIGINAL PAPER

Negotiating becoming: a Nietzschean critique of large language models

Simon W. S. Fischer[1] · **Bas de Boer**[2]

Published online: 19 June 2024

Abstract
Large language models (LLMs) structure the linguistic landscape by reflecting certain beliefs and assumptions. In this paper, we address the risk of people unthinkingly adopting and being determined by the values or worldviews embedded in LLMs. We provide a Nietzschean critique of LLMs and, based on the concept of will to power, consider LLMs as will-to-power organisations. This allows us to conceptualise the interaction between self and LLMs as power struggles, which we understand as negotiation. Currently, the invisibility and incomprehensibility of LLMs make it difficult, if not impossible, to engage in such negotiations. This bears the danger that LLMs make reality increasingly homogeneous by recycling beliefs and creating feedback loops that ultimately freeze power struggles and thus consolidate the status quo. In view of this, LLMs constrain self-formation. Based on our critique, we provide some recommendations on how to develop interactions with LLMs that enable negotiations that allow for different ways of being

Keywords Large language models · Nietzsche · Self-formation · Negotiation

Introduction

Generative deep learning models have become prevalent in recent years and especially in recent months. The most well-known models belong to the GPT family (Generative Pre-trained Transformer), like ChatGPT. These large language models (LLMs) can perform various natural language processing tasks, such as writing articles, summarising and translating texts, answering questions, structuring search engine results (Metzler et al., 2021), producing text-based games and computer code (Dale, 2021, p. 115), and generating images in combination with other models (Patashnik et al., 2021). A key feature of LLMs is that they generate other artefacts, which in turn are used by both humans and machines to create even more artefacts, such as texts for knowledge production. All of those generated artefacts relate to previous assumptions and beliefs that are reflected in the training data. As such, OpenAI, a major developer of generative deep learning models, notes in a recent report that:

> AI systems will have even greater potential to reinforce entire ideologies, worldviews, truths and untruths, and to cement them or lock them in, foreclosing future contestation, reflection, and improvement (OpenAI, 2023, p. 9).

This quote points to the danger of people unthinkingly adopting certain assumptions contained in the output generated by LLMs. In other words, the world, which is partly shaped by the technologies we have developed, acts back on us and also shapes us (Willis, 2006). In this paper, we therefore propose a way of relating to LLMs that allows to mitigate this increasingly automated creation of ourselves.

The main concept in our proposal is that of *self-formation*. As we will show, the work of Friedrich Nietzsche proves helpful in conceptualising the self and its process of becoming. Based on Nietzsche's ontology of *will to power*, the self is understood as a will-to-power organisation. That is, the self is a dynamic and relational being that is interwoven with other entities, including technologies. Due to its lack of a predetermined essence, the self is formed through interactions with others, which can be understood as *power*

✉ Simon W. S. Fischer
simon.fischer@donders.ru.nl

[1] Department of AI, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

[2] Department of Philosophy, University of Twente, Enschede, The Netherlands

struggles that take the form of negotiating boundaries set by the social context in which the self is situated (Aydin, 2021).

For self-formation, indeterminacy is particularly important because it helps to promote plurality and ambiguity, enabling the self to develop in a variety of directions. Our hypothesis is that LLMs can also be understood as will-to-power organisations with which the self interacts, but which potentially limit the self's indeterminacy and thus self-formation. To show how LLMs put constraints on self-formation, we first outline our Nietzschean approach to self-formation. Second, we frame LLMs as will-to-power organisations that structure reality and show why it is difficult to negotiate the generated output, which limits the indeterminacy of the self. Third, we identify several ways in which the development of LLMs can be improved so that negotiation becomes possible. Finally, we emphasise that LLMs are embedded in a larger socio-political context consisting of multiple will-to-power organisations, suggesting that developing LLMs differently is only a first step towards the possibility of more deliberate self-formation.

An interactionist view of self

Recall the problem mentioned above: by shaping the linguistic landscape, LLMs have the potential to reinforce ideologies, worldviews, and truths and untruths. We suggest that analysing how LLMs do so requires to focus on how they shape the process of self-formation. We propose an interactionist view of the self that conceptualises the self as a dynamic and relational being that is formed through interactions with other entities, including technologies (Aydin, 2021). In this section, we will introduce relevant concepts from Friedrich Nietzsche to explain self-formation.

Nietzsche's will to power ontology

While we will not give a full account of Nietzsche's ontology (e.g., Richardson, 1996), we want to show the difference to the Cartesian understanding, which regards the self as an independent, invariable and static entity that is decoupled from its environment (e.g., Birhane, 2021, p. 3).

For Nietzsche, the self is characterised by its lack of essence. He understands the human being as the *as-yet-undetermined animal* (Nietzsche, 1996, §62), whose nature is undefined, just as its development remains unfinished and its ends unknown. The self is, in other words, not a stable entity with a predetermined essence, but inherently indeterminate. This understanding can be linked to a broader ontology, which assumes that

> *unless something happens, there is nothing at all*. This not only means that events are ontologically prior to what is, but also that being is derived from events rather than the other way around (Aydin, 2021, p. 36).

Strictly speaking, the self never *is*, but becomes through *inter*-actions that happen *between* the self and others (Kyselo, 2014, p. 8), with a multitude of possible and therefore uncertain outcomes. The self is ultimately understood as a variable, dynamic and relational entity that is situated within its (technological) environment. It is the environment, however, that limits the indeterminacy of the self, because it embodies particular ideals and values that lead to conventions and expectations.

To emphasise the relational and indeterminate aspect, Nietzsche's conceptualisation of the self must be considered in the light of his ontology of *will to power*. For Nietzsche, 'all reality is will to power' (Aydin, 2021, p. 42). Even negating this proposition or trying to resist the 'game' of will to power is an act of power (Aydin, 2007, p. 26). It is thus impossible to step out of the game of power. Power has two important characteristics. First, 'power is only power in relation to another power' (Aydin, 2021, p. 42). Consequently, power is not static and independent, but inherently relational and thus dynamic. Second, power always strives for more power (Aydin, 2021, p. 43). Hence, power has no pre-determined end, but is by nature indeterminate. The relationality and indeterminacy of power require *organisation*, *struggle* and *negotiation*.

Organisation, struggle, and negotiation

The will to power must not be understood as a single force. Instead, there is a multiplicity of wills to power. The self is a will to power among (and not separate from) other wills to power, which is why reality is 'a permanent chaos at work' (Aydin, 2007, p. 27). For the self to form a unity and not fall apart, this chaos must be organised. Accordingly, the self is a diverse organisation that is the result of organising or integrating different power relations through interaction. So instead of considering all reality as will to power, a more fitting description would be 'all reality is 'will to power' organizations' (Aydin, 2007, p. 30). Thereby, the seemingly 'stable essence' or identity is a temporary projection that is subject to change. Put differently, with a well-organised will-to-power organisation, the illusion of stability and independence occurs (Aydin, 2021, p. 46). A will-to-power organisation can, for example, organise itself as a student, parent, or friend, depending on the context in which it is situated. The dynamic will-to-power organisation is therefore what it is through the interactions with other will-to-power organisations.

Organising oneself while embedded in different power relations is an act of struggle. Struggle is not understood as an act of violence or dominance, but rather as growth

(Aydin, 2007, p. 40). Growth, in turn, is understood as the self transcending its current state. Nietzsche (2006, §4) illustrates this in the following analogy:

> What is great about human beings is that they are a bridge and not a purpose: what is lovable about human beings is that they are a *crossing over* and a *going under*.

The self is like a bridge, *between* two states, the actual and the possible, which also illustrates its dynamic nature. In the *crossing over* the self transcends from one state to the other. For this, the self must overcome itself; it is the *going under* of the actual that constitutes the self. In other words, the self becomes by overcoming its present state, which is an ongoing process because there is no end to power. To grow, the self must be able to organise its struggle with other will-to-power organisations. To do so, it must first organise the tension it senses (e.g., opposing ideals) in order to channel it effectively in a directed manner. Tension, in other words, must be organised. Hence, struggle and organisation are interlinked and mutually dependent.

Struggle can be understood as negotiation. The self discusses, questions or contests 'actual' and present ideals, values and identities with other will-to-power organisations to bring about the 'possible'. We distinguish between the interactive process of negotiating meaning, which requires understanding, and a procedural negotiation, i.e., the process of how to reach an agreement, which we will discuss briefly in Sect. 4. Overall, negotiation ideally fosters the growth and flourishing of the self, which we call *self-formation*.

Given its indeterminacy, namely its lack of a pre-established essence, the self can never be fully grasped or defined either by others or by itself. The self is never identical with the image that others have of it, while the self never coincides with itself. Accordingly, the negotiation of others' representations and the urge (i.e., will to power) to overcome its lack (i.e., indeterminacy) enables self-formation.

In view of this, the self is neither in full control, nor a mere plaything. The self is defined by others, but also defines itself. Accordingly, a tension arises between the determinacy and indeterminacy of the self, whereby neither of these two poles should predominate too much. This tension can be captured by the distinction between 'hetero-formation' or 'patient-constitution', a formation imposed by others, and critical self-formation, as an activity of the self. If the self were completely heteronomously determined, this would undermine its agency. Either because the tension or chaos (i.e., the determining force) is too great, so that the self is unable to organise itself, or because any chaos and thus struggle is cancelled out by the affirmation and maintenance of the status quo. If the self were completely indeterminate, it could not form a unity because it would lack identifying boundaries and would disintegrate. The situatedness of the will-to-power organisation, however, always sets certain boundaries, so that self-formation is not completely arbitrary. Critical self-formation therefore rests on the ability and necessity to constantly re-negotiate and thereby overcome set boundaries, which makes self-formation not a only an individual but also a political project.

Technological self-formation

Nowadays, machine learning models are assigning identities to the self in more and more situations. In doing so, they co-determine its social class, its opportunities and its access to (material) resources (Benjamin, 2019; Eubanks, 2019). Recommender systems, for example, suggest which films to watch, which items to buy or which political party to vote for. Other systems decide who is invited for a job interview, who is eligible for a loan, or who is likely to commit a crime. Although these representations embedded in different models do not attempt to capture the totality of the self, they ultimately shape choices and actions (Verbeek, 2004), and thus co-produce the self on the basis of assumed characteristics. As such, there is a higher tendency to associate a criminal identity with categories such as 'Black' and 'Hispanic' (Mohler et al., 2018, p. 2457); or Amazon's cameras for monitoring delivery drivers that report errors even when other drivers are at fault (Gurley, 2021), thus defining and producing a 'bad' driver. These representations have far-reaching consequences. As Nietzsche (1974, §58) notes: 'what things *are called* is incomparably more important than what they are'.

Given the ubiquity of technologies with which the self interacts, 'self-formation [can be] increasingly captured as technological self-formation' (Aydin, 2021, p. 210).[1] In this process of technological self-formation, technologies are neither purely instrumental, nor completely deterministic. The self can act, but is also acted upon. Technology and self mutually constitute each other.

Self-formation in the era of large language models

At this point, we want to frame LLMs as will-to-power organisations, which allows us to conceptualise the interaction between self and LLMs in terms of power struggles and thus negotiation. In doing so, we do not want to attribute any conscious will to LLMs. Instead, we understand LLMs as the bearer or extension of other will-to-power organisations resulting from socio-technical practices, e.g., OpenAI or

[1] Technological self-formation must not be understood as self-enhancement. For a discussion, see e.g., Aydin (2017).

the worldview reflected in the data sources like Reddit. For simplicity, however, we will refer to LLMs as will-to-power organisations that reflect certain values and assumptions. Due to their widespread use in various contexts, LLMs generate outputs with these embedded values in an automated, systematic and accelerated way.

This is significant insofar as language categorises and structures reality. Describing the self as a 'citizen', 'user' or 'customer', for example, entails different expectations and assumptions (see Mooney & Evans, 2015, p. 32), shaping the way in which the self relates to its environment, as well as how it understands itself (Coeckelbergh, 2018, pp. 1506–8). Importantly, as we will show, it becomes increasingly difficult for the self to negotiate the meaning of synthetic texts. Negotiation is crucial because, as we understand it, it bridges the gap between over-reliance on machines, which leads to hetero-formation, and self-formation, which presupposes a degree of control or agency. In the following, we will show however how LLMs intensify the tension between determinacy and indeterminacy of the self.

So far, we have treated the self as a generic entity. When it comes to the interaction with LLMs, however, there are many different selves to consider: miners and manufacturers who construct the material infrastructure, data subjects whose data have been scraped (without their consent) to be fed into the models, engineers who build the model, click workers who flag inappropriate output, developers who use LLMs to build other applications, journalists or creators who produce content, and people who consume the generated content.

We will mainly focus on will-to-power organisations (i.e., selves) that consume generated content, such as news articles. In this, the self might not interact directly with the LLM, but through other services, such as a chat bot, news site, or a search engine, that make use of a language model. In a way, the LLM takes the place of the otherwise human interlocutor.

In the following we will argue (1) that the static representation of language and the self contrasts with the dynamic nature of both, which in turn reduces ambiguity, diversity and pluralism; (2) that the invisibility and incomprehensibility of LLMs undermine deliberate, reflective and reciprocal interaction in which negotiation is possible; and (3) that LLMs re-cycle old beliefs which entrench current power structures, i.e., the status quo. And since it is difficult or impossible to negotiate (2) the static assumptions (1) a reinforcing feedback loop (3) emerges. This altogether hinders (radical) change and undermines the indeterminacy of the self.

Static and reductionist representation

Of language

To understand the meaning of words (i.e., signifier) that contain domain-specific concepts (i.e., signified), we need to consider context (Mitchell, 2020, p. 226): Who says something, in which situation, with what kind of intonation and with what intention. Hence, not only quantitative, but also qualitative aspects are relevant, which are based on an implicit shared understanding and tacit knowledge and which resolve (or reduce) the ambiguity of language (e.g., Bisk et al., 2020). The word 'light', for example, can refer to physical weight, the intensity of colour, or to a source of illumination. Meaning is thus not reducible to words, making language inherently contextual and reciprocal. Language, in other words, cannot be decoupled from use.

In natural language processing, word embeddings encode (to a certain extent) and represent the meaning of words in relation to other words. In this process, words are reduced to numerical values and probabilistic correlations. Put differently, a language model is a probability distribution over a sequence of words that correlates the occurrence of words (Bender & Koller, 2020). For this, data determine the epistemic boundaries of LLMs, and it is often assumeed that more data leads to 'better' results (boyd danah & Crawford, 2012, p. 663). Although models can optimise the target function that links input to output by processing more data, the quantity of data does not automatically lead to diverse word sequences and associated viewpoints (see Bender et al., 2021, p. 613).[2]

For GPT-3, 60% (equivalent to 570 GB) of the training data comes from the web scraping repository Common Crawl (Brown et al., 2020, pp. 8–9). An initial analysis of this sheer volume of data indicates that it 'contains a significant amount of undesirable content'(Luccioni & Viviano, 2021). Another large part (22%) comes from WebText2 dataset, which collects data from Reddit, where demographics are primarily associated with white young men (Bender et al., 2021, p. 613). This results in certain views and assumptions being over-represented, known as representation or sampling bias (Blodgett et al., 2020, p. 5455).

[2] GPT-4 might contain more contextual information compared to previous models, as more computing resources are available to enable the processing of larger amounts of data. Nevertheless, we do not assume that more computing power and more data could solve the problem of static and reductionist representation. This is because the dynamic and context-dependent (i.e., qualitative) character of language cannot be (adequately) quantified. Whether models can derive a sufficient representation of the world from data alone is an ongoing debate (e.g., Mitchell, 2020, p. 267; Winograd, 1990, p. 179).

Data is always a partial representation of the phenomenon, in this case natural language. Hence, the representation of a phenomenon is highly normative and an act of power, because it is a matter of judging what is considered desirable, adequate or 'normal' (Gururangan et al., 2022; Miceli et al., 2020; West, 2020). This judgement can be implicit (Boon, 2020), as the data used to train machine learning models often reflects the views of engineers (Raji et al., 2021, p. 8), who are a relatively small and homogeneous group in themselves with a particular value system (Cave & Dihal, 2020; Denton et al., 2021; Crawford & Paglen, 2019).

It may not be the goal to adequately represent the entirety of language with LLMs, but the danger lies in their broad application across various contexts. By perpetuating the beliefs of the 'speakers' reflected in the training data, the synthetic texts generated by LLMs convey a comparatively narrow worldview and ultimately a particular, and in a sense, 'normalised' or 'standardised' understanding of the self.

Of the self

In view of the above, several studies suggest that LLMs generate narratives that contain harmful and stereotypical beliefs. For example, GPT-3 creates texts in which women are more associated with emotions and family and are portrayed as less powerful, while men are associated with politics, sports or war (Lucy & Bamman, 2021, p. 50). Next, women are associated with occupations such as nurse, receptionist or housekeeper, while men are linked with occupations associated with a higher level of education, such as banker or professor (Brown et al., 2020, p. 36). Furthermore, the 'Black' race is associated with low sentiment (Brown et al., 2020, p. 37). Similarly, a sentiment classifier (not using GPT-3) rates the sentence 'Let's go get Chinese food' lower than 'Let's go get Italian food' (Speer, 2017). Next to that, GPT-3 associates religious groups such as Jews with money and Muslims with terrorism or violence (Abid et al., 2021; Brown et al., 2020, p. 38). It also makes undesirable connections in the context of disability, such as linking mental illness with homelessness (Hutchinson et al., 2020). In summary, the synthetic texts generated by GPT-3 convey particular and harmful assumptions in relation to gender, race, religion and ableism. Crucially, those who are most vulnerable to these assumptions embedded in LLMs are already underprivileged or marginalised within society. The widespread use of LLMs thus reinforces and intensifies existing power structures (e.g., Birhane 2021; O'Neil 2016).

Instead of arguing that 'bias' should be removed, we want to stress that LLMs preserve certain values, forming a sedimented horizon of meaning to which the self must relate in one way or the other, thereby co-constituting the self. In other words, the static representations embedded in LLMs co-produce the self. Although LLMs may be less static at a technical level compared to previous approaches, they remain static on a conceptual level in that they organise reality in a certain way and thereby charge the self with categories and values. Forsythe(1993) describes how knowledge is (not) embedded in early expert systems, which is still prevalent in current LLMs:

> In everyday life, the beliefs held by individuals are modified through negotiation with other individuals; as ideas and expectations are expressed in action, they are also modified in relation to contextual factors. But the information encoded in a knowledge base is not modified in this way. (p. 466)

Again, it may not be the goal to (adequately) represent or categorise the self with LLMs. Besides, the self does not have to identify with or agree with the generated output, but can question it. Accordingly, alternative interpretations and thus ways of being remain (theoretically) possible. A survey of 963 Facebook users from 2019, for example, shows that 260 participants (27%) disagree with the labels assigned to them by the platform, while 491 participants (51%) feel uncomfortable being categorised (Hitlin & Rainie, 2019, p. 7). The dissatisfaction indicates that other interpretations remain possible. Nevertheless, the confrontation with assumed characteristics and the non-acceptance thereof forms the self and subsequent interactions in a certain way that is highly individual. Importantly, although 568 out of the 963 participants (59%) think that Facebook categorises them correctly, they do not know why this is the case, which brings us to our next point.

Unidirectional interaction

As mentioned, there are various selves or will-to-power organisations that interact with LLMs. Some might be aware of the implications of LLMs, but lack the know-how or means to act accordingly (as an individual), while others may not even know about the existence of LLMs. The question then arises as to who has the power to challenge and negotiate the values embedded in LLMs. Although the following will not lead to a breakdown of different stakeholders, we want to argue that the invisibility and incomprehensibility of LLMs undermine the possibility to modify and negotiate the embedded worldview.

Invisibility of LLMs

The self cannot escape the fact that it is always defined by others. Other people also categorise the self based on behaviour and assign a (static) identity to it accordingly. This is, however, likely to happen in a shared and situated context. The self is thus involved in practices of meaning-making in which it has some agency of self-representation. Besides,

the reciprocity and (theoretically) open-ended nature of an interaction or dialogue between people facilitates questioning, contesting or refuting certain views, which allows for re-interpretation and re-evaluation of assumptions. These reciprocal interactions are less rigid or static leaving room for (contextual) ambiguity and negotiation. A dialogue is, in other words, (more or less) indeterminate and thereby accounts for the fluidity and plurality of the self.

This informed, reflective and contextual interaction or dialogue with LLMs is very difficult, if not impossible. For various machine learning models, the self may not realise that it is interacting with them and, more importantly, what this interaction entails. American students, for example, were unaware that their Facebook feed was algorithmically sorted and filtered according to presumed interests (Powers, 2017). For GPT-3, participants in a study could not tell whether news articles of about 500 words were written by a human or a machine (Brown et al., 2020, pp. 25–26). So, unless disclosed, the self may not be aware that it is confronted with generated content that might be tailored to its interests.[3] Although this might raise concerns about trust and responsibility, the point is that the seamless nature of LLMs makes them invisible actors in the process of self-formation.

Both invisibility[4] and unawareness lead to an illusion of independence, which creates a false impression of autonomy of the self, over which LLMs supposedly have no effect. If the self is, however, unaware of who or what it is interacting with, power is not absent, but concealed due to its inherent relationality.[5] While the same might be true for ideologies and cultural values that operate in the background, it is in principle possible to negotiate the meaning of these values with the people who adhere to them. LLMs, on the other hand, perpetuate and consolidate the already pervasive influence of culture, and by concealing the interaction and the reasons why certain outputs were generated, the self is prevented from (effectively) organising the power struggle, as it cannot (meaningfully) interact with the originator. The interaction with these models is therefore not reciprocal, in which two entities form each other, but unidirectional, in which one entity is a passive recipient of the power exercised by the other.

A will-to-power organisation that interacts intentionally with LLMs can, of course, modify the generated text, translation or summary. As studies show, however, machines can influence word choice (Brandstetter & Bartneck, 2017, p. 284), and writing-assistants can influence the opinions of authors (Jakesch et al., 2023), and even corrupt moral judgement (Krügel et al., 2023). This should be of concern as people tend to over-rely on machines (Buçinca et al., 2021). Further, it becomes more difficult to modify the output in cases where LLMs process search engine queries, or are implemented in other applications like chat bots.

Moreover, if the self changes the output this does not (immediately) prevent LLMs from continuing to produce similar narratives (see Sect. 3.3). It is possible to adjust LLMs through reinforcement learning from human feedback (RLHF), but this is a time-consuming, labour-intensive and exploitative process (Perrigo, 2023), and also raises the question of whose values the model should be 'aligned' with (e.g., Manders-Huits, 2011). Besides, given the high computational and environmental costs of re-training a LLM, only the weights of the network are updated so that the 'base model' (provisionally) remains the same. Implemented safeguard filters can therefore often be circumvented by slightly adjusted prompts (often referred to as 'jailbreaking'). Even though this may seem like a dynamic and reciprocal interaction, LLMs do not fundamentally and promptly change with regard to the underlying static assumptions. Put differently, LLMs have an influence on the self, but rarely, if ever, does the self influence LLMs.

Incomprehensibility of LLMs

The complexity of LLMs leads to uncertainties about how words are encoded in the model. Despite great efforts in explainable and interpretable AI, the problem of understanding why LLMs generate a certain output remains. If explanations do provide insights, which can sometimes be misleading (Rudin, 2019), we are reminded that machine learning models make correlations that do not necessarily represent the world in a meaningful way. A medical chat bot run by GPT-3, for example, suggested starting to recycle or even committing suicide to overcome sadness (Rousseau et al., 2020). Similarly, Meta's LLM called Galactica, which was shut down after a few days of its release, generated an article on the health benefits of eating crushed glass.

Although we may not need to comprehend how the model arrives at such (nonsensical) conclusions in order to reject the generated output, the inherent incomprehensibility of LLMs undermines the anticipation of harmful consequences, shifting away forward-looking responsibility from the companies developing and deploying these systems. In the case

[3] Current efforts to watermark synthetic texts would allow to disclose the use of LLMs. The effectiveness of this technique has yet to be proven.

[4] Machine learning models are highly dependent on physical resources, so they are not invisible in the sense that they are immaterial.

[5] In *Discipline and Punish: The Birth of the Prison* (1995), Michel Foucault argues that disciplinary power, most notably in prisons, but also in 'factories, schools, barracks, hospitals, which all resemble prisons' (p. 228), is used to alter behaviour and correct individuals on the basis of what is considered 'normal'. The prisoner becomes their own guard by internalising the discipline so that the power goes unnoticed.

of a medical chat bot suggesting suicide, it is questionable whether a (mentally unstable) patient should be responsible for judging whether the suggestion is appropriate or not. Besides, even if it were the case that a therapist suggested suicide, the patient could turn their attention to the therapist and respond, organise its struggle and direct its tension specifically towards the cause.

Furthermore, while these inadequate (and harmful) correlations were detected, we may not (yet) be aware of the more subtle correlations of word occurrences embedded in LLMs (e.g., Blodgett et al., 2020, p. 5460). Also because language is context-dependent and concepts, values or other prejudices vary (Weidinger et al., 2021, p. 12). Moreover, concepts are often deeply rooted and taken for granted. When imagining a 'surgeon', for example, one might think of a man (Coeckelbergh, 2018, p. 1508), or when referring to '*women doctors* as if *doctor* itself entails not-woman'(Bender et al., 2021, p. 617). It is thus not surprising that LLMs reflect such power-related gender stereotypes (Kotek, 2023).

The invisibility and incomprehensibility of GPT undermine the possibility to question and challenge embedded assumptions, resulting in the self not being able to co-construct and (effectively) negotiate the meaning generated (e.g., De Jaegher & Di Paolo, 2007). The self becomes a patient in a unidirectional interaction in which LLMs form the self, but in which it is difficult, almost impossible, for the (individual) self to shape LLMs. Given this unidirectionality, the interaction between self and LLMs does not enable self-formation, but can be better understood in terms of hetero-formation.

Reinforcing the status quo

When a will-to-power organisation (i.e., LLM) is no longer contested, it is recognised as truth, and eventually becomes reality (Aydin, 2007, p. 36). While a certain degree of stability is required to live life, some stable organisations or 'truths' can also be harmful. Besides, the fact that power always strives for more power already implies some sort of homogenisation. Homogenisation, however, is not desirable, for it cancels out the tension or chaos of the will-to-power organisation. Homogenisation or uniformity, in other words, prevents change. Similarly, LLMs prevent change, as they neither establish a new hierarchical order, nor collapse an old one.

Machine learning models use historic data to make future predictions. In cases where the problem space is limited and well explored, such as for chess or skin cancer detection in medical images,[6] this static relation between past and future might be appropriate and even necessary. For 'the model is operating within a background of existing scientific understanding'(Sullivan, 2019, p. 20) and the relationship between cause and effect is unlikely to change. In view of the inherently dynamic, contextual, unfinished and thus unpredictable nature of both language and self, however, this (scientific) certainty or stability is not given.

Again, LLMs do not attempt to predict the self or the totality of language, but the static representation nonetheless leads to recycling particular beliefs, such as assuming that a family consists (exclusively) of a married woman and man with children (Weidinger et al., 2021, p. 13). The problem is that the synthetic data generated by LLMs and other generative models are almost identical to the input data (e.g., Abid et al., 2021). Another project named 'This person does not exist', for example, generates images of faces that resemble to a high degree the faces in the training data (Webster et al., 2021). By generating ever more similar data, synthetic data can contaminate future training sets (Brown et al., 2020, p. 29) and thus increase the confirmation bias of LLMs.

The same might be said about the self; ideals or language do not emerge out of nowhere. The self, however, is more fluid and more likely to change. Machine learning models, in contrast, do not (sufficiently) account for temporal changes and are thus more rigid or static. To learn new correlations, the model requires thousands of examples that may not yet be available (Weidinger et al., 2021, p. 12). Hence, '[L]LMs become increasingly outdated with time' (Lazaridou et al., 2021, p. 9). So although LLMs are more dynamic compared to simple if-then algorithms and can be adjusted by human feedback (reinforcement learning), the underlying problem remains that LLMs create a static representational meaning that is not easily changed and, importantly, often outdated. LLMs therefore remain static on a conceptual level by organising language and thus reality on the basis of past data, thereby reinforcing the status quo.

Feedback loops

Perhaps it can be argued that LLMs allow the self to challenge the status quo by exposing the current power structures and beliefs in certain communities. This, in turn, would allow the self to contrast them to other ideals and overcome them accordingly. This apparent 'advantage', however, should not be used to justify the many negative consequences of LLMs. While LLMs may enable desirable self-formation for some, they are detrimental to the self-formation of many others, especially those already marginalised by society.

By prioritising and reproducing certain narratives over others, feedback loops emerge (O'Neil, 2016) and with them the danger of 'value-lock' (Bender et al., 2021, p. 614). For predictive policing, for example, a crime is predicted based

[6] And even these systems can still have limitations and can result in poor performance with darker skin tones.

on historic data. In the case of an arrest, data is created that is fed back to the system, confirming the prediction. If no arrest is made, no data is generated, hence there is no information to correct the false assumption of the model. This leads to an increase of police patrol in certain areas (Mohler et al., 2018).

Feedback ultimately optimises the predetermined target function and thus performance of the model. But as Rosenblueth et al. (1943, p. 19) note, '[a]ll purposeful behaviour may be considered to require negative feed-back'. The lack of feedback, in other words, does not alter the design of the system (Franklin, 1990, p. 49). Rather, it promotes the characteristic of power, namely that power always strives for more power. As a result, these models create an environment that justifies the initial assumptions.

Importantly, feedback loops also apply to 'good' ideals the self finds desirable. This is because the self tends to believe things that are similar to what it already believes (Mansoury et al., 2020). With LLMs it would be possible to produce personalised stories with other specific (behavioural) data points, creating so-called 'filter bubbles'. The self might think that it is acting according to its own will, while in reality it takes the ideals it has adopted from others for granted and reinforces them. The problem with 'own evaluations' and 'opinions' is, as Nietzsche (1997) states:

> what they do is done for the phantom of their ego which has formed itself in the heads of those around them and has been communicated to them; - as a consequence they all of them dwell in a fog of impersonal, semi-personal opinions (§105).

Since the self is a social and self-made product, 'own' choices or ideals are certainly never entirely one's own. Personalised stories, however, make it increasingly difficult to question currently held ideals. Moreover, the way language transports ideals can itself be very subtle, such as word order that maintains power hierarchies (Mooney & Evans, 2015, p. 112). By perpetuating past behaviour, LLMs create a feedback loop, for themselves and for the self that makes it increasingly difficult to contrast values and assumptions with other alternatives (e.g., Weidinger et al., 2021, p. 14). In doing so, LLMs neither establishe a new hierarchical order, nor do they collapse an old one, but maintain and consolidate current power structures (Blodgett et al., 2020; Birhane et al., 2021). For Nietzsche, the blocking of new forms of life through the freezing of power struggles is 'life threatening' (Aydin, 2021, p. 173).

The self is undoubtedly a historical being that is susceptible to various behavioural or dispositional 'feedback loops'. There is a difference, however, between being determined by a past and re-enacting a past. While the former is imposed by others (e.g., machine learning models), the latter is a more reflective and conscious process. Namely, to the extent that the self can negotiate and re-appropriate the meaning of its own past (see Haste, 2004, p. 414). Nietzsche calls this 'active forgetting' (Aydin, 2017). Active forgetting does not mean simply ignoring or erasing the past, but rather overcoming certain ideals or events (i.e., to grow) by re-interpreting and re-negotiating them. So in contrast to LLMs, which reiterate the past by assuming that the future resembles the past, the self projects itself towards its future through transcendence and overcoming.

LLMs therefore limit self-formation in two ways. First, a will-to-power organisation that identifies with the values contained in the generated output does not sense any tension because the views are in agreement and taken for granted. The will-to-power organisation can maintain its stable unity. But by affirming currently held beliefs, chaos and continuous struggle are increasingly suppressed. Accordingly, the self does not transcend its current state. Second, a will-to-power organisation that disagrees with these values is likely to feel a strong tension. But the invisibility and incomprehensibility of LLMs undermine the possibility of an organised struggle or negotiation over the generated output. In other words, the will-to-power organisation is unable to channel its tension effectively towards the cause.

Chaos and alternative worlds

Constant struggle to overcome the current state is certainly not pleasant as it also requires taking responsibility. Hence, the fact that LLMs contribute to some form of uniformity might sound reassuring (to some). From a Nietzschean point of view, however, uniformity resists the

> permanent chaos at work, which is a condition for discovering evermore and alternative worlds. The chaos is, therefore, not a mere burden that we have to overcome to survive or make our life easier; that is only one aspect of it. It also plays a very positive role. It is the basis for all creation and creativity. Without it, nothing novel could emerge. The more that chaos breaks into our ordered world, the more our creative power is stimulated (Aydin, 2021, pp. 43–44).

Nietzsche highlights the importance of chaos by characterising two types of will-to-power organisations, namely the strong or healthy and the weak or sick (Alfano, 2015; Aydin, 2007, p. 39). The strong type is characterised by being well organised (i.e., a seemingly stable self), while at the same time possessing an intense tension or chaos that may stem from opposing ideals or desires, or from a recognition of one's indeterminacy. The greater this tension, the stronger the will-to-power, but the easier it is for the the will-to-power organisation to fall apart. If the chaos is too great, which the self cannot organise or channel, this is a sign of weakness for Nietzsche. The same is true if the will-to-power organisation

has no tension or chaos to resolve, because without struggle or constant re-negotiation there is no growth. Therefore, in order to grow (i.e., having the urge to overcome the status quo), the challenge is to maintain a balance between being well organised (i.e., a seemingly stable unity) and to sense a tension or chaos that can be channelled effectively.

We consider that the strong will-to-power organisation aims at self-formation. For this we regard the indeterminacy of the self as an intrinsic value. Instead of conforming to imposed or 'standardised' ideals, i.e., the status quo (e.g., Fromm, 2006, pp. 150–1), the self liberates and forms itself through ongoing re-evaluation and re-negotiation (see Aydin, 2021, p. 173). In doing so, the self has the agency or control to deliberately relate to and create its own ideals and goals (e.g., Haste, 2004, p. 426), while still always running the risk of becoming a weak will-to-power organisation that possesses no chaos and becomes, in a sense, powerless.

A sick or weak self or the will-to-power organisation becomes the patient of the other will-to-power organisations, and is determined by other forces (i.e., 'hetero-formation'). Either because the self is not able to organise the tension and thus the struggle to which it is subjected, or because all tension is reduced and thus struggle is excluded. In both cases, the indeterminate nature of the self is ultimately undermined. While the self may no longer be indeterminate in the sense of undefined, it remains indeterminate in the sense of unfinished. It is, however, unlikely that the self, while engaged in subsequent power struggles in the process of becoming, will pursue its 'own' desired or defined goals.

It could be argued that by setting its own ideals and goals, the self undermines its own indeterminacy. But these goals or definitions will always be acknowledged or challenged by others (see Alfano, 2015, p. 266). As mentioned earlier, power is only power in relation to other power. The general assumption nevertheless is that 'self-determination' or 'self-realisation' through negotiation promotes the flourishing of the self and ultimately facilitates a 'good' and meaningful life. The ambiguity of a 'good' life emphasises the importance of overcoming 'normalised' ideals and instead valuing diversity and plurality, i.e., chaos, to allow for alternative and multiple ways of being (e.g., Escobar, 2018). Accordingly, indeterminacy can also be an instrumental value, such as to promote an open and pluralistic society, which we leave open for further research.

Recommendations

To enable different ways of being, we need to create opportunities to counter and resist the 'uniform value system' that LLMs and their outputs create. To this end, negotiation should be the focal point. To allow for negotiation, however, it is necessary to address more than just the final meaning of an output (i.e., backward-looking). Rather, the entire process of model construction and validation must be negotiable in order to create desirable conditions in the first place (i.e., forward-looking). Put differently, negotiation can take place in at least three different phases, namely model construction, model validation and model interaction.

We want to stress that we should not fall prey to an implicit technological solutionism and, in particular, to instrumentalism, which assumes that once all limitations are solved, LLMs will function as mere tools. Instead, LLMs alter our relations to the world and our experiences of it, ultimately constituting our being. Nonetheless, we can reduce the constraints on self-formation by improving our interactions with LLMs, aiming to increase human agency and self-determination.

During model construction, we need clear practices and rules for data and model documentation for the respective purpose (Mitchell et al., 2019). This includes making transparent which data was collected, how it was cleaned and annotated (e.g., Gebru et al., 2020; Jo & Gebru, 2018) and why it was valued as useful (Plasek, 2016, p. 6). Accordingly, the appropriateness and usefulness of the model can be questioned (Stilgoe, 2023). In addition, audits and review mechanisms enable to anticipate the consequences of potential errors (Raji et al., 2020). Model validation is particularly important as evaluative benchmarks tend to generalise the performance of the model (Raji et al., 2021), especially since training data and test data increasingly overlap. Accordingly, contextual domain knowledge and 'non-expert' participation (Birhane et al., 2022) is becoming increasingly important, provided that participation also empowers the community concerned (Sloane et al., 2020).

During model use, we need to design reflective and deliberate interactions that are not unidirectional, but reciprocal. At the very least, as currently stipulated in the European AI Act, this means disclosing the use of a LLM, which could result in the self refusing the interaction in the first place (i.e., algorithmic aversion), or even trusting it more (i.e., automation bias). Next, it must be possible to provide (negative) feedback by modifying or deleting data points or by rejecting data collection altogether.

In order to avoid over-reliance and thus hetero-formation, design choices that highlight the uncertainty of the model could further dismantle the notion of an all-knowing machine. An answer of a LLM could, for example, be preceded with 'I assume', although this could at the same time increase the risk of further anthropomorphisation. Another solution could be to rethink our interactions with machine learning models in general. Instead of presenting options for the self to choose from, we could delegate the decision-making from the machine to the self from the outset, for example, by stimulating critical reflection and increasing

human engagement through machine-generated questions (Haselager et al., 2023).

Concluding remarks

The starting-point of our analysis was that the self is formed through interactions in which it is always confronted with ideals or assigned identities. Due to its indeterminacy, however, the self can never be fully defined, neither by itself nor by others. Self-formation is thus a constant re-negotiation of values or ideals that enables the self to overcome its current state and to grow and (radically) change in the process.

As we suggested, the static representation of language embedded in LLMs bears the danger of eliminating ambiguity and plurality. In addition, negotiating the meaning of synthetic texts is difficult if not impossible. As a result, ideals and assumptions become more and more uniform and homogeneous, excluding alternative and new ways of being (e.g., Bianchi et al., 2022). Accordingly, chaos and struggle that are necessary for growth are suppressed, weakening the will-to-power organisation in the process of self-formation.

To reduce the constraints of self-formation, we therefore need different ways of interacting with LLMs. As we suggest with our approach, the focus should be on negotiation during model construction, validation, and interaction. As LLMs enable different services and applications, they affect different practices, whether social, political, or economic in varying degrees. Further research is necessary that considers the lived experiences of selves with a particular application applied to a particular context (see Blodgett et al., 2020, p. 5458). At the same time, the self bears the responsibility for deciding which power struggles or interactions it wants to participate in and how. After all, LLMs are not the only will-to-power organisations that form the self, rather there are various other (non-technical) will-to-power organisations. We therefore do not claim that LLMs completely debilitate indeterminate self-formation per se.

Nevertheless, given their increasing proliferation, LLMs have an impact on society at large, beyond the immediate interaction with a particular self, by shaping the linguistic landscape in which we live. The question of how LLMs will organise society remains open, and no particular development is prescribed. In view of our technological culture, however, with the overarching acceptance and pursuit of quantification, categorisation, and prediction, one could say that machine learning systems attempt to organise our lives in a uniform way, just as Christianity did according to Nietzsche. And in doing so, LLMs reduce chaos and indeterminacy. Our greater concern is thus the extent to which this uniformity runs counter to the ideal of an open and pluralistic society that promotes different ways of being. Especially as LLMs, as bearer of values of other will-to-power organisations (e.g., data sources like Reddit, OpenAI), consolidate current power structures. We should therefore be aware that focusing on the 'optimal' functioning of LLMs is not the final solution. The necessary solutions are not (primarily) technical, but political, concerning the organisation of the social realm in which deliberate self-formation can take place. In the end, self-formation is not only an individual, but also a political project.

Acknowledgements We would like to thank Ciano Aydin for crucial feedback at an early stage, and the Societal Implications of AI & CNS group at the Donders Institute for comments on a presentation of an earlier version the paper. We also thank Mark Alfano and an anonymous reviewer for their valuable remarks.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* pp. 298–306. https://doi.org/10.1145/3461702.3462624

Alfano, M. (2015). How one becomes what one is called: On the relation between traits and trait-terms in nietzsche. *The Journal of Nietzsche Studies, 46*(2), 261–269. https://doi.org/10.5325/jnietstud.46.2.0261

Aydin, C. (2007). Nietzsche on reality as will to power: Toward an organization–struggle" model. *Journal of Nietzsche Studies, 33*, 25–48. https://www.jstor.org/stable/20717895.

Aydin, C. (2017). The posthuman as hollow idol: A nietzschean critique of human enhancement. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 42*(3), 304–327. https://doi.org/10.1093/jmp/jhx002

Aydin, C. (2021). Extimate technology: Self-formation in a technological world (first). *Routledge*. https://doi.org/10.4324/9781003139409

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pp. 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new jim code*. Polity.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2022). November 7. *Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale*. arXiv:2211.03759 [cs].

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2). https://doi.org/10.1016/j.patter.2021.100205

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*.https://doi.org/10.1145/3551624.3555290

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). The values encoded in machine learning research. https://arxiv.org/abs/2106.15590

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Boon, M. (2020). The role of disciplinary perspectives in an epistemology of scientific models. *European Journal for Philosophy of Science, 10*(3), 31. https://doi.org/10.1007/s13194-020-00295-9

boyd danah, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5). https://doi.org/10.1080/1369118X.2012.678878

Brandstetter, J., & Bartneck, C. (2017). Robots will dominate the use of our language. *Adaptive Behaviour*, *25*(6). https://doi.org/10.1177/1059712317731606

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–21. https://doi.org/10.1145/3449287

Cave, S., & Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology, 33*, 685–703. https://doi.org/10.1007/s13347-020-00415-6

Coeckelbergh, M. (2018). Technology games: Using wittgenstein for understanding and evaluating technology. *Science and Engineering Ethics, 24*(5), 1503–1519. https://doi.org/10.1007/s11948-017-9953-8

Crawford, K., & Paglen, T. (2019). September 19. *Excavating AI: The politics of training sets for machine learning*. https://excavating.ai

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113–118. https://doi.org/10.1017/S1351324920000601

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences, 6*, 485–507. https://doi.org/10.1007/s11097-007-9076-9

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society, 8*(2), 1–14. https://doi.org/10.1177/20539517211035955

Escobar, A. (2018). *Designs for the pluriverse: Radical interdependence, autonomy, and the making of worlds*. Duke University Press.

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police and punish the poor*. St Martin's Press.

Forsythe, D. E. (1993). Engineering knowledge: The construction of knowledge in artificial intelligence. *Social Studies of Science*, *23*(3), 445–477. http://www.jstor.org/stable/370256

Franklin, U. (1990). *The real world of technology*. CBC Enterprises.

Fromm, E. (2006). *Escape from freedom* [Die furcht vor der freiheit] (L. Mickel & E. Mickel, Trans.; 13th ed.). Deutscher Taschenbuch Verlag. (Original work published 1941)

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., & Crawford, K. (2018). Datasheets for datasets. *CoRR*, *abs/1803.09010*. http://arxiv.org/abs/1803.09010

Gurley, L. K. (2021). September 20. *Amazon's ai cameras are punishing drivers for mistakes they didn't make*. Retrieved October 10, 2023, from https://www.vice.com/en/article/88npjv/amazons-ai-cameras-are-punishing-drivers-for-mistakes-they-didnt-make

Gururangan, S., Card, D., Dreier, S. K., Gade, E. K., Wang, L. Z., Wang, Z., Zettlemoyer, L., & Smith, N. A. (2022). Whose language counts as high quality? measuring language ideologies in text data selection. https://arxiv.org/abs/2201.10474

Haselager, P., Schraffenberger, H., Thill, S., Fischer, S., Lanillos, P., Van De Groes, S., & Van Hooff, M. (2023). Reflection Machines: Supporting Effective Human Oversight Over Medical Decision Support Systems. *Cambridge Quarterly of Healthcare Ethics, 1–10*,. https://doi.org/10.1017/S0963180122000718

Haste, H. (2004). Constructing the citizen. *Political Psychology, 25*(3), 413–439. https://doi.org/10.1111/j.1467-9221.2004.00378.x

Hitlin, P., & Rainie, L. (2019). Facebook algorithms and personal data. *Pew Research Center*.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in nlp models as barriers for persons with disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. https://doi.org/10.1145/3544548.3581196

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 306–316. https://doi.org/10.1145/3351095.3372829

Kotek, H. (2023). *Chatgpt doubles down on gender stereotypes even when they don't make sense in context*. https://twitter.com/HadasKotek/status/1648453764117041152

Krügel, S., Ostermaier, A., & Uhl, M. (2023). The moral authority of ChatGPT. https://arxiv.org/abs/2301.07098

Kyselo, M. (2014). The body social: An enactive approach to the self. *Frontiers in Psychology, 5*, 986. https://doi.org/10.3389/fpsyg.2014.00986

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Ruder, S., Yogatama, D., Cao, K., Kocisky, T., Young, S., & Blunsom, P. (2021). Mind the gap: Assessing temporal generalization in neural language models. https://arxiv.org/abs/2102.01951v2

Luccioni, A., & Viviano, J. (2021). What's in the box? an analysis of undesirable content in the common crawl corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

Natural Language Processing, pp. 182–189. https://doi.org/10.18653/v1/2021.acl-short.24

Lucy, L., & Bamman, D. (2021). Gender and representation bias in gpt-3 generated stories. *Proceedings of the Third Workshop on Narrative Understanding*, pp. 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

Manders-Huits, N. (2011). What Values in Design? The Challenge of Incorporating Moral Values into Design. *Science and Engineering Ethics, 17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2145–2148. https://doi.org/10.1145/3340531.3412152

Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum, 55*(1), 1–27. https://doi.org/10.1145/3476415.3476428

Miceli, M., Schuessler, M., & Yang, T. (2020). Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction, 4*, 1–25. https://doi.org/10.1145/3415186

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency,* pp. 220–229. https://doi.org/10.1145/3287560.3287596

Mitchell, M. (2020). *Artificial intelligence: A guide for thinking humans*. Pelican.

Mohler, G., Raje, R., Carter, J., Valasik, M., & Brantingham, J. (2018). A penalized likelihood method for balancing accuracy and fairness in predictive policing. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2454–2459. https://doi.org/10.1109/SMC.2018.00421

Mooney, A., & Evans, B. (2015). *Language, society and power: An introduction* (4th ed.). Routledge. https://doi.org/10.4324/9781315733524

Nietzsche, F. (1966). *Beyond good and evil: Prelude to a philosophy of the future* (W. Kaufmann, Trans.). Vintage Books.

Nietzsche, F. (1974). *The gay science: With a prelude in rhymes and an appendix of songs* (W. Kaufmann, Trans.). Vintage Books.

Nietzsche, F. (1997). *Daybreak: Thoughts on the prejudices of morality* (M. Clark & B. Leiter, Eds.; R. J. Hollingdale, Trans.). Cambridge University Press.

Nietzsche, F. (2006). *Thus spoke zarathustra* (A. Del Caro & R. Pippin, Eds.; A. Del Caro, Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9780511812095

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

OpenAI. (2023). March 14 *GPT-4 Technical Report*.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2085–2094.

Perrigo, B. (2023). January 23 *Openai used kenyan workers on less than \$2 per hour to make chatgpt less toxic*. Time. Retrieved September 7, 2023, from https://time.com/6247678/openai-chatgpt-kenya-workers/

Plasek, A. (2016). On the cruelty of really writing a history of machine learning. *IEEE Annals of the History of Computing, 38*(4), 6–8. https://doi.org/10.1109/MAHC.2016.43

Powers, E. (2017). My news feed is filtered? *Digital Journalism, 5*(10), 1315–1335. https://doi.org/10.1080/21670811.2017.1286943

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. *35th Conference on Neural Information Processing System (NeurIPS 2021)*.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. arXiv:2001.00973 [cs] . http://arxiv.org/abs/2001.00973

Richardson, J. (1996). *Nietzsche's System* (1st ed.). Oxford University Press. https://doi.org/10.1093/0195098463.001.0001

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10*(1), 18–24. https://www.jstor.org/stable/184878

Rousseau, A.-L., Baudelaire, C., & Riera, K. (2020) October 27 *Doctor GPT-3: Hype or reality?* Nabla. Retrieved January 5, 2022, from https://www.nabla.com/blog/gpt-3/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. http://arxiv.org/abs/1811.10154

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. https://doi.org/10.48550/ARXIV.2007.02423

Speer, R. (2017). July 13 *How to make a racist ai without really trying*. Retrieved November 16, 2021, from https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Stilgoe, J. (2023). We need a Weizenbaum test for AI. *Science*, *381*(6658), eadk0176. https://doi.org/10.1126/science.adk0176

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035

Verbeek, P.-P. (2004). *What things do: Philosophical reflections on technology, agency, and design*. Pennsylvania State University Press.

Webster, R., Rabin, J., Simon, L., & Jurie, F. (2021). This person (probably) exists: Identity membership attacks against GAN generated faces. https://arxiv.org/abs/2107.06018

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., & Gabriel, I. (2021). Ethical and social risks of harm from language models. https://arxiv.org/abs/2112.04359

West, S. M. (2020). Redistribution and rekognition: A feminist critique of algorithmic fairness. *Catalyst: Feminism, Theory, Technoscience, 6*(2), 1–24.

Willis, A.-M. (2006). Ontological Designing. *Design Philosophy Papers, 4*(2), 69–92. https://doi.org/10.2752/144871306X13966268131514

Winograd, T. (1990). Thinking machines: Can there be? are we? In D. Partridge & Y. Wilks (Eds.), *The foundations of artificial intelligence* (pp. 167–189). Cambridge University Press.