The Indeterminate Self and Large Language Models

A Nietzschean Critique of GPT-3

A Master's Thesis by

Simon Fischer

22nd March 2022

Supervisor: Dr. Bas de Boer
Second Reader: Prof. Dr. Ciano Aydin

MSc Philosophy of Science, Technology and Society
Faculty of Behavioural, Management, and Social Sciences
University of Twente
Enschede, the Netherlands

One must still have chaos in oneself
to be able to give birth to a dancing star.
I say to you:
you still have chaos in yourself.

Friedrich Nietzsche

Abstract

The advent of large language models introduces a new relation between the self, language and technology. Language enables interaction through which the self experiences and understands itself in relation to others. As GPT-3 generates synthetic texts that convey ideals, it co-constitutes these interactions. The self, language, and GPT-3 are ultimately intertwined.

To conceptualise this relation and anticipate challenges, a new understanding of the self is required. For traditional notions of the self, such as the essentialist and dualist, assume the self to be static, independent and invariable. This, however, sustains the impression that technology has no effect on the self. And it also excludes the possibility of (radical) change.

Nietzsche's will to power ontology, in contrast, conceives of the self as having no pre-established essence. Instead, the self is a priori undefined and its development remains unfinished and unknown. The self is thus inherently indeterminate and always in a state of becoming. In this, it is formed through interaction. Becoming is not an autonomous process, however, as the self is always constrained to some degree by the social context in which it is embedded.

Self-formation thus requires constant re-evaluation and re-negotiation of set boundaries. Given the indeterminate nature of the self, negotiation is particularly important. For otherwise, the self would conform to 'normalised' ideals, which in turn denies different ways of being. The overarching goal, then, is to secure the indeterminacy of the self, by allowing for ambiguity and pluralism.

GPT-3, however, debilitates self-formation because (1) the static representation of language conveys pre-determined identities of the self, negating the ambiguity and plurality of both; (2) the invisibility and incomprehensibility of GPT-3 undermine deliberate and reflective interaction in which negotiation is possible; (3) GPT-3 re-cycles old assumptions that reinforce the status quo. And as the negotiation of the static assumptions is undermined, GPT-3 creates a self-reinforcing feedback loop. This ultimately excludes other alternatives and hinders (radical) change.

Keywords: self-formation, will to power, negotiation, ambiguity, change

Acknowledgements

Philosophy is said to be therapeutic. I can certainly say that about my thesis. It was was challenging, insightful, self-revealing and overall rewarding. I want to express my gratitude to those who have formed this work and thus also me. Thanks to Jochem Zwier for the initial supervision and support in setting the direction of my thesis; for bringing me closer to Nietzsche's philosophy; for guiding me through the realm of arts; and for your patience while I was busy with work. Many thanks to Bas de Boer for taking over the supervision after Jochem left the UT and for all your efforts in this task. Thank you for making me realise that my initial interest in art and creativity was not the ideal use case; for our inspiring dialogues and your countless and challenging comments, which ultimately helped me to organise the chaos; and also for all your kind words that went beyond the thesis project. Thanks to Ciano Aydin for providing crucial feedback, the scope of which I only understood much later. And well, thanks for your engaging theory that got me thinking over the last months – and will likely continue to do so. Thanks to Koray Karaca for our discussions, not only for the thesis, but also for various paper topics during the programme, which are in a way part of this work. I thank each of you for listening and I hope I have listened well enough. Thanks also to the many others who were indirectly involved – including Baba Haft, because like Nietzsche said, without music, life would be a mistake. Without the help of all of you, this thesis would not have been possible (which also shows that the self is not an independent entity). Finally, thank you, the reader, for considering reading the following (or just these lines).

Contents

Abstract **ii**

Acknowledgements **iv**

List of Figures **viii**

Introduction **1**

Brief Literature Review 2

Theoretical Framework 3

Research Question and Structure 7

Hypothesis 8

1 The Indeterminate Self **10**

1.1 The Static and Independent Self 11

1.1.1 Aristotle's Essentialism 11

1.1.2 Descartes' Mind-Body Dualism 12

1.1.3 The Inadequacy of the Essentialist and Dualist Views . 14

1.2 The Dynamic and Relational Self 16

1.2.1 Nietzsche's Will-to-Power Ontology 17

1.3 Technological Self-Formation 21

1.3.1 Prerequisites for Self-Formation 23

1.3.2 Why the Indeterminate Self is Desirable 25

1.4 Conclusion 26

2 The Workings and Worldview of GPT-3 **28**
2.1 What is GPT-3? . . . 29
2.2 A Brief History of Machine Learning . . . 30
2.2.1 Imitation of Intelligence . . . 31
2.2.2 From Rules . . . 32
2.2.3 To Patterns . . . 33
2.3 The Functioning of GPT-3 . . . 36
2.3.1 Generative Model . . . 37
2.3.2 Task-Agnostic Performance . . . 40
2.4 The Importance of Data Sets . . . 42
2.4.1 The 'Normalised' Self . . . 43
2.5 Conclusion . . . 46

3 Language and Self-Formation **48**
3.1 Language Games and Forms of Life . . . 49
3.1.1 Public Language . . . 51
3.1.2 Social Categories . . . 53
3.2 Conclusion . . . 54

4 Self-Formation in the Era of Large Language Models **55**
4.1 Tension between Determinacy and Indeterminacy . . . 56
4.2 The Challenges that GPT-3 poses for Self-Formation . . . 58
4.2.1 Static and Reductionist Representation . . . 60
4.2.2 Unidirectional Interaction . . . 63
4.2.3 Reinforcing the Status Quo . . . 68
4.2.4 Uniformity . . . 74
4.3 Conclusion . . . 77

Conclusion **80**
Limitations . . . 80
Outlook . . . 81
Summary . . . 83

Bibliography **86**

List of Figures

2.1 Discriminative and Generative models 38
2.2 A simplified encoder-decoder architecture 40
2.3 Zero-shot, one-shot and few-shot task settings 41

3.1 Language use, language games and form of life 50
3.2 Saussure's model of the sign 51

Introduction

In recent years, large language models (LLMs) have become very popular, both in scientific research and in the public interest. This is because these models are able to perform tasks that were previously unimaginable made possible through the increase in computational resources and the amount of available data. Especially the language model developed by the company OpenAI and released in May 2020 raised a lot of media attention.

The so-called Generative Pre-trained Transformer 3 (Brown et al., 2020), in short GPT-3, is able to write articles, summarise and translate texts, produce text-based games as well as computer code (Dale, 2021, p. 115). It can be used to answer questions in the form of chat bots or to structure search engine results (Metzler et al., 2021). And in combination with other models, GPT-3 can generate images (Patashnik et al., 2021; Ramesh et al., 2021; Reed et al., 2016). Overall, GPT-3 can perform various natural language processing tasks.[1]

GPT-3 does so by processing the instruction it is provided. The Guardian, for example, published an article titled 'A robot wrote this entire article. Are you scared yet, human?'. In the editor's note it states that:

> For this essay, GPT-3 was given these instructions: 'Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI.' (GPT-3, 2020)

[1]For some examples of GPT-3 see https://gpt3demo.com/.

Simply put, GPT-3 works like a large auto-complete function. To generate synthetic text, GPT-3 was trained on a large amount of data from which it derived patterns of word occurrences and thus learned a linguistic representation. I will elaborate on this in chapter 2.

At the time of release, GPT-3 was by far the largest language model.[2] The size of the model is relevant in that it relates to the quality of the output.[3] GPT-3, thus, performs better than its predecessors. Given the ability to accomplish various tasks combined with the (in some cases) rather impressive quality of the output, GPT-3 is both fascinating and disturbing. In the following, I will briefly discuss various speculations and concerns.

Brief Literature Review

In the case of the Guardian article the nature of the author was disclosed. For some texts, however, it is difficult to distinguish whether it was written by GPT-3 or a human. In a test study with 62 participants, the mean accuracy in correctly attributing the author of news articles of about 500 words was 52%, which is equivalent to guessing (Brown et al., 2020, pp. 25–26). In other words, the participants could not tell whether the article was written by a human or by GPT-3. As a result, GPT-3 has revived the question of the Turing Test, namely whether machines can think (Elkins & Chun, 2020). Many highlight the fact, however, that regardless of how convincing the output might be, GPT-3 has no real understanding of meaning (Bender & Koller, 2020; Marcus & Davis, 2020). Others nevertheless raise concerns that GPT-3 could make authors and journalists obsolete (Floridi & Chiri-

[2]With its 175 billion parameters, GPT-3 is about hundred times bigger than its predecessor. Google's latest language model Switch-C, released in January 2021, consists of 1.6 trillion parameters (Fedus et al., 2021). GPT-3, however, appears to be better at performing a broader range of tasks, which is demonstrated by various examples available.

[3]A small language model can still perform well on a specific task.

atti, 2020, p. 691). As the editor of the Guardian article also notes, '[e]diting GPT-3's op-ed was no different to editing a human op-ed' (GPT-3, 2020). In view of this, the ease of text production and dissemination leads to another likely consequence, namely the spread of misinformation and thus radicalisation (McGuffie & Newhouse, 2020). In order to counterbalance and limit the misuse and harmful societal consequences of GPT-3, OpenAI controls the access to its service (Brockman et al., 2020).[4] Microsoft, however, has an exclusive license for GPT-3 (Scott, 2020) and has only recently introduced its first products that use the language model to 'help users build apps without needing to know how to write computer code or formulas' (Langston, 2021). And considering Microsoft's involvement, it is only likely that the proliferation of GPT-3 will increase. Similarly, researchers at Stanford University claim that:

> AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. (Bommasani et al., 2021)

Theoretical Framework

Within the field of philosophy of technology, one of the main objectives is to conceptualise the relation between humans and technology, and how technology mediates experiences and practices (Ihde, 1990). This is significant because the way technology, especially machine learning models, perceive and present the world to us is increasingly becoming our dominant reality. Think of how we follow navigation systems, Netflix recommends which movie to watch, Fitbit motivates us to run or political microtargeting influences elec-

[4]Until November 2021, GPT-3 was not available to the public. Instead, one was put on a waiting list after application. Now, this waitlist has been removed as 'wider availability made possible by safety progress' (OpenAI, 2021).

tions. As these examples illustrate, technology affects the way we navigate in the world, shaping our experiences and thinking (Malafouris, 2021), as well as our moral decisions (Verbeek, 2004). Since GPT-3 generates language, the basis of human communication, it is reasonable to assume that it will also shape our beliefs, values, experiences and actions.

In this process, we should understand technology as neither completely deterministic nor purely instrumental. On the one hand, this is because technological determinism understands Technology as a uniform and autonomous force that structures society. In doing so, it neglects the fact that there are various technologies, as well as other social and political forces. The instrumentalist view, on the other hand, sees technology as something that does not affect the human or society, but merely as a tool that the human uses at will. In doing so, it attributes too much autonomy to the individual and overlooks the fact that these seemingly neutral tools themselves shape human decisions and actions. In the end, both views are inadequate, which is why we need a new concept for the relation between humans and technology.

In his recent book, *Extimate Technology: Self-Formation in a Technological World*, Ciano Aydin (2021) provides a third perspective besides the instrumentalist or deterministic perspective of technology, namely the *interactionist* approach.[5] It acknowledges that technology neither has no influence on the human nor does technology completely control the human. Rather, humans and technology shape each other reciprocally. This also means that the interactionist view does not understand the self (i.e., human being) as a static or fixed entitiy independent of its environment. Instead, the self and technology are interconnected, technology is part of the self.

This dynamic and relational understanding of the self contrasts with a particular and influential view in the history of Western philosophy. Namely,

[5]There are also other theories that fall into the interactionist approach (see Aydin, 2021, p. 6). But most of them do not manage to 'overcome the categorical separation of human beings and technology' (Aydin, 2021, p. 5).

that of Aristotle and René Descartes. I will discuss both views in more detail in section 1.1. But in short, both regard the self as static, invariable and independent (Aydin, 2021, p. 31). This is, however, problematic. Because if we understand the self as independent of its (technological) environment, we fall back to the instrumentalist viewpoint, with the assumption that the self is completely autonomous and that technology has no influence on the self. And if we understand the self as static and invariable, we tend towards the deterministic view in which the self is denied the possibility of (radical) change and is instead subject to external forces. Aristotle's and Descartes' understanding of the self is thus not sufficient to conceptualise the reciprocal shaping of self and technology, including potential challenges in this process.

The interactionist self provided by Aydin (2021) moves beyond Aristotle's essentialist and Descartes' dualist self. A central concept for the interactionist self Friedrich Nietzsche's *will to power* ontology. I will discuss this in section 1.2. But Nietzsche does not consider the self as a pre-established and invariable entity. Instead, the self is a priori undefined and its development remains unfinished and unknown. The self is thus understood as inherently indeterminate. This means, the self never 'is', but is always in the process of becoming (Aydin, 2021, pp. 42, 56; Nietzsche, 1967, p. 280). Thereby,

> the direction of the development of reality is not a priori established, thereby guaranteeing the possibility of fundamental novelty and radical change; the world is, in a sense, continuously pregnant with a measureless variety and multiplicity of possibilities that are still unknown to us. (Aydin, 2021, pp. 50–1)

In the process of becoming, the self is formed through interactions (i.e., power struggles) with others, including technology. Accordingly, the self is a relational being. This, however, already indicates that becoming is neither a completely autonomous nor a purely arbitrary process. Instead, the self is always constrained and determined to some degree by society and the (technological) environment. That is, the self is always embedded in a context in which certain ideals and assumptions pre-exist. It is then through identific-

ation with and negotiation of these ideals through which the self 'becomes' and forms itself, i.e., self-formation.

As the environment becomes more saturated with technology, it also increasingly transports and conveys certain ideals and assumptions. This is especially true for large language models such as GPT-3. Because language makes it possible to express and exchange beliefs or goals. This also means, language constitutes the actions of the self, and is at the same time the precondition for interaction. Accordingly, the self experiences and understands itself (in relation to others) through language. The self, language and GPT-3 are thus intertwined and co-constitute each other.

This is relevant because as Clowes et al. (2021) notes:

> Our creation of artefacts to better shape the world to our own purposes has reciprocally transformed the nature of the human cognitive profile. We have recreated ourselves in the image of our tools. (p. 15)

With large language models, we not only have created an artefact, but one that is itself capable of creating artefacts (i.e., synthetic texts that convey ideals), which in turn both the self and GPT-3 use to create further artefacts, and so on. This raises the question of how the the self can control and negotiate the increasingly automated 'recreation of itself' by an increasingly complex and incomprehensible technology. Espescially considering that language is contextual and thus ambiguous, which means it is, like the self, never completely fixed. Accordingly, language and the ideals it carries are fluid and change over time. LLMs, however, strive for order and regularity by formalising language and predicting the probability of future word occurrences based on historic data. This, then, raises the further question of how the self can 're'-create itself with a technology that assumes the future resembles the past.

In the end, this leads to a tension. On the one hand, language and the self are both fluid, ambiguous, open-ended and thus indeterminate. On the other hand, GPT-3 reduces language to numerical values, re-cycles old

beliefs, and thus renders language, including assumptions, static. And while the self negotiates ideals in dialogue with others in a shared and situated context, GPT-3 systematically transports ideals across contexts trough its widespread use.

Research Question and Structure

This brings me to my research question of my thesis:

> *How does the interaction with synthetic texts generated by large language models like GPT-3 debilitate indeterminate self-formation?*

As mentioned, self-formation is always constrained by others, since the self is not independent. In other words, the self is always in a state of tension between self-formation (i.e., indeterminacy) and 'hetero-formation' (i.e., determinacy). So my assumption is that GPT-3 intensifies this tension and thus debilitates indeterminate self-formation.

To answer my research question, I have to address several sub-questions first, which provide the structure of the thesis. These questions are:

1. How is the self understood? What role does technology play in self-formation? What are the prerequisites for self-formation? Why should an indeterminate self be sustained?

2. How does GPT-3 learn a linguistic representation? What values, ideals and views of the self do the generated texts convey? Why is this problematic?

3. What role does language play in self-formation?

In chapter 1, I will provide a theoretical understanding of the self. To do so, I will first discuss Aristotle's essentialism and Descartes' dualism in order to contrast it with Nietzsche's will to power ontology. This allows me to frame self-formation as technological self-formation. Moreover, I will formulate

necessary conditions for self-formation and point out why an indeterminate self is desirable. In chapter 2, I will elucidate the workings of GPT-3. To better contextualise the language model and to demystify its capabilities I will present a brief history of machine learning. I will then show that GPT-3 is not neutral or objective, but contains certain values based on the data it has been trained on. And by transporting these values across contexts, it 'normalises' ideals and the view of the self. In chapter 3, I will briefly turn to Ludwig Wittgenstein and the concept of language games to show that language and use are interwoven. Language thus affects how the self relates to its environment and understands itself accordingly. In the final chapter 4, I will bring together the findings and address my research question. A conclusion including a brief outlook follows thereafter.

Hypothesis

The question of whether and how GPT-3 debilitates self-formation has a normative aspect. Namely, it presupposes necessary conditions for self-formation as well as assumptions as to why an indeterminate self is desirable. As such, I understand self-formation as a deliberate, reflective and intentional process, that enables the self to overcome its past by negotiating certain ideals or assigned identities. Given the indeterminate nature of the self, negotiation is particularly important. For otherwise, the self would conform to 'normalised' or 'standardised' ideals, which in turn denies different ways of being. The overarching goal, then, is to secure the indeterminacy of the self, by allowing for ambiguity and pluralism. This ultimately allows the self to change in ways that it finds desirable and that go beyond defined categories captured by numerical values.

At the same time, however, complete indeterminacy is not desirable. For the self needs identifying boundaries to which it can relate to in order to form a unity. But since the self is embedded in a context and thus not independent, certain boundaries are always given. Self-formation, then, is

the constant re-negotiation of those boundaries.

In view of this, I will argue that GPT-3 debilitates indeterminate self-formation for the following reasons. First, the ambiguity of language is reduced to a static linguistic representation, resulting in pre-established and preconceived ideals, beliefs and identities of the self. Second, the self cannot negotiate these derived meanings. Because the interaction between the self and GPT-3 is not reciprocal, but unidirectional. Given the invisibility of GPT-3, the self is often unaware that it interacts with the system in the first place. In addition, built-in assumptions that lead to an output are increasingly difficult to question or challenge due to the incomprehensibility of the model. Third, synthetic (and behavioural) data is generated during the interaction and fed back to the model. This feedback loop affirms the static representations, especially since they cannot be negotiated. As a result, GPT-3 reinforces the status quo, which opposes change.

Despite these challenges, GPT-3 does not render self-formation of a particular self impossible. Simply because GPT-3 is not the only force that forms the self. Nevertheless, given the widespread use and increasing tendency of LLMs, ideals on a societal level are also likely to be affected, which in turn shape the self. Thus, self-formation is also a social endeavour.

I will leave possible solutions for further research. But overcoming the challenges is ultimately a social endeavour. Many machine learning systems fail in real world settings (D'Amour et al., 2020), so they should not be seen as universal solutions. Instead, the importance of the context in which these systems are used must be recognised. This also means acknowledging that each data set represents a certain world view, hence merely 'unbiasing' data is insufficient. Instead, these models, including their construction, need to be open to investigation and negotiation to a set of diverse people. Finally, interactions and processes are required in which the self can negotiate and modify the meaning of the generated output.

Chapter 1

The Indeterminate Self

This chapter lays the theoretical foundation of my thesis. I will develop the notion of the self as an inherently indeterminate being, formed in the process by the interactions with others, including technology.

The question that guides the first part is: *How is the self understood.* Influential concepts of the self in Western philosophy are Aristotle's essentialism, and Descartes' mind-body dualism.[1] I will first discuss both views and then show why they are insufficient to understand the interaction between self and technology.

So to conceptualise self-formation, a new ontological understanding of the self is necessary. For this I will build on the framework of the *interactionist self* by Ciano Aydin (2021). The interactionist self is influenced by theories from Charles S. Peirce, Friedrich Nietzsche, Jacques Lacan and Sigmund Freud. I will primarily focus on Nietzsche and his concept of *will to power*. This is because the will to power ontology frames the self as a dynamic and relational being, which allows me to describe the interaction between the self and technology. Accordingly, in section 1.3, I will address the question of *what role does technology play in self-formation.*

[1]The following is not an exhaustive account or comparison of the various conceptions of the self (e.g., Gallagher, 2011).

This is crucial because for Nietzsche the self has no pre-established essence, but is the result of its interactions. In other words, the self is inherently indeterminate and in the process of becoming. But despite its indeterminacy, the context in which the self is embedded sets certain boundaries. Accordingly, I will formulate necessary conditions for self-formation in section 1.3.1, which are primarily characterised by negotiation. This will help me in chapter 4 to examine how LLMs challenge these conditions. But before that, in section 1.3.2 I will point out why an indeterminate self is desirable.

1.1 The Static and Independent Self

Aristotle and Descartes provide two views of the self that have been influential in Western philosophy. I will discuss both briefly and show why they are inadequate to understand the interaction between the self and technology.

1.1.1 Aristotle's Essentialism

For Aristotle, every being and also every object has an essence. It is 'the what it is to be' (Cohen and Reeve, 2021) or 'the "whatness" of a given entity' (Fuss, 1989, p. xi). The essence of something highlights similarities of entities that belong to the same category, and thus also the differences from entities that do not belong to that category. As such, for Aristotle, the essence of human beings, in contrast to other animals, is rational and virtuous behaviour. Thereby, the essence that defines the thing is pre-established and inscribed in the thing (Aydin, 2021, p. 24).

This essentialist thinking leads to the assumption that the self can be characterised or categorised by different attributes. For example, by gender, race or ethnicity. But similar aspects might be attributed to different selves, and at the same time these aspects can vary from self to self. For example, the self can be defined by (biological) sex, but not all women or men share the same (socially constructed) gender characteristics. As we see, the essence is difficult to describe. Yet it is what makes the self unique. In this sense,

essentialism can be understood in more psychological terms. The essence constitutes the 'I-ness', the identity, of the self.

Importantly, for Aristotle, the essence does not change (Bickhard, 2009, p. 548). This means that although traits, preferences or appearance of the self are likely to change, the essence nevertheless remains the same. The essence, i.e., the identity, is thus understood as something static, invariable and stable (Aydin, 2021, p. 24).

The Potential to Change

Although the essence remains unchanging, Aristotle acknowledges that the self changes. Thus, he introduced the notion of *potentiality*. That is, everything has the potential to change (Aydin, 2021, p. 33). But for Aristotle everything has a goal or *telos*, known as the teleological structure (Aydin, 2021, p. 23). The telos is already inscribed in the essence, or rather, '[t]he essence of a thing is, at the same time, its goal' (Aydin, 2021, p. 34). The telos of an acorn, for example, is to become an oak tree. For the self, by defining and attributing it an essence, a goal is also prescribed, which is linked to ideals and assumptions. The self is thus assumed to act in a certain way because it has characteristic x. I will resume to this shortly. In the end, the teleological structure presupposes that change follows a predetermined path.

1.1.2 Descartes' Mind-Body Dualism

Objective Knowledge

Renés Descartes, who is widely considered the founder of modern philosophy, wanted to create a basis for objective knowledge. Thereby his view of the self was highly informed by the technology of the camera obscura (Aydin & de Boer, 2020, pp. 731–2). In the camera obscura, light falls through a hole, projecting an object on one side onto a surface on the opposite side. Based on this understanding, for Descartes, the self perceives the *outer* world through its senses, which is then experienced or processed by the *inner* realm.

As Descartes however realises, sense perception can be sometimes deceptive. This is the case with hallucinations or optical illusions. For example, a spoon in a glass of water appears bent, even if it is not. Hence, Descartes does not trust his senses. In addition, he considers the possibility that an evil demon is the source of his erroneous beliefs, which he cannot rule out. Consequently, he also rejects all his former beliefs.[2] As he notes:

> I realized how many false opinions I had accepted as true from childhood onwards, and that, whatever I had since built on such shaky foundations, could only be highly doubtful. (Descartes, 2008, p. 13)

So to provide a foundation for objective truth, Descartes (2008) starts his First Meditation by questioning hitherto knowledge (see Husserl, 1970, p. 76).

Inside-Outside Distinction

The consequence of Descartes' epistemological inquiry to objective knowledge is that it introduces a distinction between material and immaterial substances (Aydin, 2021, p. 25). The material entails the external world including the body, while the immaterial is constituted by rational thought, i.e, the mind.

Similar to the camera obscura, for the self, information travels from the outside (i.e., material) to the inside (i.e., immaterial). Sense perception may influence or deceive thinking, but it can still be doubted by pure reason, ideally leading to freedom from erroneous beliefs (Aydin, 2021, p. 31). It is the ability to think and to doubt, to be a 'thinking thing', which for Descartes constitutes the stable and unchanging essence of the self.

Ultimately, for Descartes, objective knowledge can only be acquired by the inside (i.e., the mind) (see Aydin & de Boer, 2020, p. 4). To do so, the immaterial has to detach itself from the material. Accordingly, the mind and the body are considered as two separate and independent substances. A

[2]Descartes method of doubt is also known as hyperbolic doubt.

consequence thereof is, as I will show, that 'his [Descartes'] understanding laid the groundwork for the symbol-processing machines of the modern age' (Winograd, 1990, p. 168).

1.1.3 The Inadequacy of the Essentialist and Dualist Views

Overall, Aristotle and Descartes consider the essence and thus the self to be static and independent. As Aydin (2021) elaborates:

> From this perspective, an essence secures that a thing can possess a basic, invariable identity, despite being subject to continuous change; it makes it possible that a thing can form a unity, despite having different parts and properties; and it enables that a thing can be separated from other things, despite being involved in various interactions. (p. 31)

This understanding introduces two problems. First, the notion of independence gives the false impression of an autonomous self. Second, despite Aristotle's teleological structure, the static and invariable essence (i.e., identity) does not allow for becoming and (radical) change.

Illusion of Independence

For Aristotle, the essence is inscribed *in* the self. This means my essence or identity is likely to be different from yours. The self is thus understood as a separate being among other separate beings. For Aristotle, it remains an unsolved puzzle how an independent self interacts with another independent self (Aydin, 2021, p. 35). And Descartes' distinction between the inside and outside reinforces the idea that the self is independent of its environment. From the essentialist perspective, it is very difficult, if not impossible, to understand the relation of beings and things.

Accordingly, this understanding does not provide space to adequately conceptualise the interaction between the self and technology, as it presents

the self as a completely autonomous being. This is crucial because technology is not a mere instrument that the self uses to realise its intentions. Instead, technology shapes choices, such as a health tracking app that nudges the self to go for a run. In doing so, technology assigns (assumed) identities and ideals to the self. I will elaborate on this in section 1.3.

Moreover, as I will show in chapter 3, the self interacts with others through language. Thereby, language use is constituted by context (i.e., the 'outside') which in turn constitutes the actions and beliefs of the self (i.e., the 'inside'). In other words, language cannot be de-coupled from use. Large language models, like many other machine learning models, however, reinforce the assumption that information, i.e., words, merely pass from the outside world and can be adequately processed by internal computation.

In the end, the inside and the outside, the self and the other, are not independent, but are in a dynamic and reciprocal relation. In other words, the self is not separate, but embedded in its (technological) environment. This also means the self is not a completely autonomous being, but it is always constrained to a certain degree (section 1.3).

Change as Predetermined

Next to the illusion of independence, the essentialist and dualist views cancel out the possibility of (radical) change. For Descartes, sensory perception does not shape the 'thinking thing'. Even if he would not deny that the material world can influence thinking, the challenge remains for him to overcome these erroneous beliefs that deceive rational thought. Thereby, the stable essence, the immaterial substance, ultimately remains the same. Similarly, Aristotle acknowledges the possibility to change, but he understands change as reaching the predetermined state, i.e., *telos*. Thus, he does not consider change as essential (Aydin, 2021, p. 33).

This static understanding implies that 'something cannot evolve into something of a completely different nature' (Aydin, 2021, p. 34). That is, once a self is defined and thus ascribed an essence or identity, certain charac-

teristics, ideals and behaviours are presupposed. The result is a prejudiced and reductionist view of the self. For example, GPT-3 generates texts that associate women with emotions and men with sports. In this sense, the pre-established essence or definition determine the self to some extent. This should not be understood as inevitable determinism. Rather, this static understanding of the self does not take into account the negotiation of assigned ideals and identities through which the self can change and become. I will elaborate on this in the following sections.

1.2 The Dynamic and Relational Self

To overcome the assumption of a pre-determined, independent and unchanging essence of the self that the essentialist and dualist views advocate, a new ontological understanding of the self is necessary. This in turn makes it possible to conceptualise the interaction between self and others, including technology. Further, prerequisites (section 1.3.1) and challenges (section 4.2) for self-formation can be formulated accordingly.

In the following I will present Ciano Aydin's *interactionist self* and with it Friedrich Nietzsche's will-to-power ontology[3] on which it is based. Unlike Aristotle and Descartes, the stable self with a pre-given essence is not taken as the starting point. Rather, the essence of the self is the result of interactions. In other words, there is not first a stable self that then enters an interaction, but the self 'is' or becomes first through the relations and interactions it has (Aydin, 2021, pp. 36, 42, 56). Thus, as Aydin (2021) notes:

> *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. (p. 36)

[3]Aydin uses the term *event ontology*.

Accordingly, the self and its essence emerge in *inter*-action, which happens *between* the self and others (including technology). To give an example: The essence or identity of the self is not prescribed by being a student, but through interaction with other selves characterised as students or professors, the self becomes a student. Consequently, the essence of the self (i.e., its identity including its ideals and behaviour) is not something that is a priori established and remains always present, but rather it is the outcome of an interaction. The self is therefore understood as a variable and dynamic entity that is situated within its environment.

1.2.1 Nietzsche's Will-to-Power Ontology

For Nietzsche, the essence of the self is that it has no essence. Instead, he regards the self as the 'not yet determined animal' (Aydin, 2021, p. 177). The nature of the self is thus undefined, just as its development is unfinished and unknown. In other words, the self is inherently indeterminate.

The underlying concept of this understanding is Nietzsche's concept of *will to power*. Power has two characteristics that stand in strong contrast to the essentialist and dualist conception of the self. 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 (i.e., telos), but is by nature indeterminate. The relationality and indeterminacy of power both require *organisation* and *struggle*.

Organisation

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). So the will to power must not be understood as a single force. Instead, there is a multiplicity of wills to power. Accordingly, the self is understood as a will to power among (and not separate from) other

wills to power. And technology is also will to power, I will come back to that later.

Given the multiplicity of wills to power and their dynamics, there is, as Nietzsche calls it, 'a permanent chaos at work' (Aydin, 2007, p. 27). In order for the self to form a unity and not fall apart, this chaos has to be organised:

> An important implication of the ontological status of the will to power is that reality is always necessarily organized to some degree. (Aydin, 2021, p. 45)

Accordingly, the self is an organisation (like a system) that is the result of organising (as activity) 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).

The self or will to power organisation is, however, neither invariable, nor is it independent from other will to power organisations. For the relational nature of power makes interaction inevitable. 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).

For example, the self can be described as student, parent and a runner. When the self is situated in a university, interacting with other selves that are characterised as students or staff, the self organises itself as student. When the self prepares food for its toddler, it is organised as a parent. And when interacting with a fitness app, the self becomes a runner.

In view of this, the boundaries of a will to power organisation are very fluid. And instead of conceiving of the stable self as starting point, as Aristotle and Descartes do, Nietzsche understands the self as the result of organisation. Accordingly, the self has no fixed or invariable and, above all, no pre-established identity.

Struggle

The act of organising through interaction is constituted by struggle. In other words, the will to power organisation (i.e., the self) is formed through power struggles (Aydin, 2021, p. 40). Thereby, power or struggle is not simply understood as bodily force, but rather as growth (Aydin, 2007, p. 40). Growth, in turn, can be understood as the self transcending its current state. Nietzsche (2006) 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*. (p. 7)

The self is like a bridge, *between* two states, the actual and the possible. 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*. The going under can be understood as critically questioning the values and beliefs that constitute the self.[4] It is through overcoming ideals that may not be meaningful that the self becomes. And since there is no end to power, overcoming is an ongoing process. Ultimately, the development of the self is indeterminate, and thus (radical) change is possible.

In this process, struggle is understood as growth, which I ultimately frame as negotiation. That is, the self negotiates 'actual' and present ideals, values and identities with other will to power organisations in order to arrive at the 'possible'. This is significant in that the self lacks a pre-established essence. Accordingly, it can never be fully or adequately grasped or defined. On the one hand, this means that the self is never identical to the image others (including technology) have of it. On the other hand, the self can never fully understand itself (on its own). Rather, the self or the will to power organisa-

[4]This might sound like Descartes' doubt. The difference, however, is that in this case the questioning of values can only take place in relation to others through interaction and not by an independent observer.

tion emerges through interaction *between* different will to power organisation. This, in turn, requires ongoing re-negotiation.

In order to grow or to become, the self must be able to organise its struggle or negotiation with other will to power organisations. For this, it must organise the tension (e.g., opposing ideals) it senses in order to release it (effectively) in a directed manner. In other words, the tension itself has to be organised (Aydin, 2007, p. 38). This leads to the fact that struggle and organisation are interlinked and mutually dependent.

Strong and Weak Will to Power Organisations

For Nietzsche, there are two types of will to power organisations, namely strong or healthy and weak or sick (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 (e.g., opposing ideals, beliefs, desires). The stronger the organisation, the greater this discrepancy and the easier it is for the the will to power organisation to fall apart. The challenge then is to maintain this discrepancy between stability (i.e., to be well organised) and instability (i.e., chaos). If the chaos is too intense and cannot be organised, that is a sign of weakness for Nietzsche. The same is true when the will to power has no tension or chaos to be released. Without struggle or constant re-negotiation, however, there is no growth. Thus, to prevent decay, ongoing struggle is necessary.

The fact that power always strives for more power already implies some sort of homogenisation.[5] In the case that a will to power organisation is no longer contested, it is accepted as ground form or truth and ultimately as reality (Aydin, 2007, p. 36). While some ground forms and some amount of stability are certainly required to live life, some stable organisations or

[5]The striving of power for more power must not be confused with a teleological goal. For even if the will to power reaches its goal or 'end', it can still be challenged by another will to power, and so on.

'truths' can also be harmful. Besides, in view of the weak and strong types, homogenisation is not desirable. For it cancels out the tension or chaos of the will to power organisation and thus prevents growth.

I will elaborate on this in chapter 4, but to give an example, various machine learning systems perpetuate old beliefs. In the case of GPT-3, synthetic texts associate Muslims with terrorism or women with less power. In doing so, GPT-3 preserves a certain kind of life (Aydin, 2007, p. 37). The outcome can be twofold. On the one hand, a will to power organisation that identifies with these beliefs does not sense any tension because the views align. The will to power organisation is able to 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. On the other hand, a will to power organisation that does not identify with these beliefs is likely to feel a strong tension. But, as I will show, the will to power organisation is not able to release its tension in a directed manner effectively to the cause. In other words, the possibility of an organised struggle or negotiation over the meaning generated is undermined by the invisibility and incomprehensibility of GPT-3.

1.3 Technological Self-Formation

The lack of an essence makes the self inherently indeterminate. This means the self is both a priori undefined and its development also remains unknown and open. The self is therefore always unfinished and in the process of becoming; it is in a state of 'always-being-on-the-way' (Aydin, 2007, p. 26).

Through interaction, the self forms itself in a desired direction by pursuing its goals and desires. But due to the relational nature of the self, these goals are often set or influenced by others, such as society, institutions, family or friends (e.g., Althusser, 1971). As Moeller and D'Ambrosio (2021) put it:

> In earlier times, identity was typically assigned by the social roles one was born into. Along with birth came not only one's gender

> but also one's tribal or ethnic identity, one's social class, one's profession, and one's religion. Identification then typically consisted in committing to the roles people found themselves in by embracing the norms and internalizing the values attached to these roles. (p. 10)

Nowadays, these identities and ideals are increasingly assigned and conveyed through many machine learning models. For example, predictive policing systems calculate the probability of a self committing a crime on the assumption that a criminal's identity can be determined from their past behaviour. Similarly, political microtargeting assigns political voter identities based on certain preferences. Other systems decide which job applicants are invited for an interview, or which job ads are displayed in the first place. Crawford and Schultz (2019) note that:

> Every month, more algorithmic and predictive technologies are being applied in domains such as healthcare, education, criminal justice, and beyond. (p. 1942)

Even in the home, the Internet of Things attempts to quantify and predict various aspects of the self (e.g., Zuboff, 2019). Although these models do not attempt to capture the totality of the self, they nevertheless ascribe a rather static or fixed identity to the self. As the self interacts more and more with and through technology, 'self-formation is increasingly captured as technological self-formation' (Aydin, 2021, p. 210). This also means that the categorical distinction between self and technology (i.e., 'inside-outside') can no longer be sustained. Instead, technology becomes part of the self.

Especially LLMs, are likely to shape ideals and beliefs of the self. This is because the self uses language to relate to its environment based on which it understands itself. For example, each framing of a relation between two selves such as consumer/producer, student/professor or doctor/patient signify different concepts that entail different ideals and expectations. Thereby, these concepts are often ingrained and taken for granted. For instance, 'referring to *women doctors* as if *doctor* itself entails not-woman' (Bender et al.,

2021, p. 617). This means, language can transport values and ideals in a very subtle manner. And through its widespread use, GPT-3 does so across contexts.

This is important because while the self is defined by others in a shared and situated context (e.g., in dialogue), it is involved in meaning-making. But since models learn through data and are based on the assumptions of engineers, the self is not (directly) involved. These models then transport a particular and, in a sense, 'normalised' or 'standardised' worldview in a systematic, automated and accelerated way.

Accordingly, in the process of (technological) self-formation, technologies (e.g., LLMs) are not mere instruments, as they are not neutral or objective, but contain certain values and assumptions. As a result, technologies mediate and shape choices and actions (Malafouris, 2021; Verbeek, 2004). At the same time, technology is also not completely deterministic in the sense that the self is a mere patient to technology. Just as the will to power is not a single force, neither is there a single or unified Technology, but different technologies. So it is uncertain how these different technologies manifest and how they in turn affect the self. So although various models assign identities to the self, its development is thus not (completely) pre-determined.

Technology has neither no effect on the self nor does it control the self entirely. Rather, 'humans and technology are seen as reciprocally and mutually shaping one another' (Aydin, 2021, p. 5). Similarly, Malafouris (2021, p. 116) argues that human thought does not occur in isolation and then unconditionally translates into action, but is interwoven with the tools one uses. In view of this, necessary conditions for self-formation can be formulated.

1.3.1 Prerequisites for Self-Formation

Self-formation must not be understood as self-enhancement. The purpose of using technology is not to become 'smarter', 'better' or 'faster' at certain activities. Besides, this would not be easy to achieve, as technologies change the very concept of what it means to be 'smart' or 'good' (Aydin, 2017). In

other words, there is no linear development to these characteristics. So instead of self-enhancement, self-formation is understood as a process in which the self can grow in a desirable way. Put differently, the self becomes what it wants to be.

The self is nevertheless always embedded in a social context. So although the self is inherently indeterminate, it is always determined or defined by others (to some extent). Be it through societal conventions, expectations, or technology allowing only limited actions. At the same time, this also means that the indeterminacy of the self does not make its development completely arbitrary. For the social context in which self-formation takes place sets boundaries to which the self relates (see Aydin, 2021, pp. 147–150).

Accordingly, the self is neither in full control nor is it a mere plaything. For this reason, Di Paolo et al. (2018, p. 7), who study the evolution and development of the body through interaction with its environment (i.e., enactivism), especially through language, argue that the self is a social and self-made product at the same time. Self-formation is therefore not a completely autonomous and individual endeavour, but also a societal challenge (Aydin, 2021, p. 116). In view of this, a distinction can be made between (indeterminate or critical) self-formation as an activity of the self and 'heteroformation', a form of formation that is imposed by others. So, forming the self and having the self formed by others.

Ultimately, self-formation is understood as a conscious and intentional process that enables the self to reflect on certain ideals or assigned identities and to identify with them accordingly. Importantly, the self can negotiate their meaning (see Haste, 2004, p. 430). In this context, negotiation through questioning, contesting or refuting ideals, is understood as power struggles. So deliberate self-formation requires the self (i.e., will to power organisation) to be able to channel its chaos or tension (e.g., opposing beliefs) in a directed manner (towards its cause). This also presupposes that the self is aware of the power struggle to which it is exposed. It may then also be that the self consciously resists or refrains from the interaction or negotiation, which

is also a form of will to power. Ultimately, the indeterminacy of the self refers to the ability to constantly re-negotiate and thereby overcome given boundaries.

If these conditions are however not met, the self or the will to power organisation becomes the patient of the opposing will to power organisation. Meaning the self 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 self becomes a weak will to power organisation and its indeterminate nature is ultimately undermined. The self may no longer be indeterminate in the sense of undefined. But it still 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.

In view of the interaction with LLMs, the self (often) does not know how the meaning of the generated output was derived. And given the invisibility and complexity of these models, the self cannot (easily) challenge these beliefs or ideals. In other words, negotiation is undermined because the self cannot effectively release its tension against the originator. This is again problematic because during the interaction between the self and the model data is generated. This data is then fed back to the system that shapes further interactions, and so on. But again, the self does not know how or what data is being processed. Accordingly, it seems as if LLMs debilitate indeterminate self-formation. I will elaborate on this in chapter 4.

1.3.2 Why the Indeterminate Self is Desirable

'Own' choices or goals are certainly never completely the self's own choices. Nevertheless, I consider the indeterminacy of the self as an intrinsic value. That is, the self has the agency or 'control' to deliberately relate to ideals and goals. Through the ability to (effectively) exercise its will to power (e.g., Haste, 2004, p. 426; Sutterlüty & Tisdall, 2019, pp. 184–5), the self liberates

itself from imposed expectations or ideals, i.e., set boundaries (e.g., Fromm, 1941/2006, pp. 150–1). So instead of conforming to the status quo and in the process becoming a weak will to power organisation, which possesses no chaos and in a sense becomes 'powerless', the self actualises and creates itself (anew) through ongoing re-evaluation and re-negotiation (see Aydin, 2021, p. 173). This, in turn, allows for ambiguity and pluralism, which does justice to the indeterminacy to the self.

The general assumption then is that this 'self-determination' or 'self-realisation' through negotiation promotes the flourishing and thus the well-being of the self, which ultimately facilitates a 'good' and meaningful life. I will not address the question of a 'good' life, since it means different things to different selves. This fact, however, underlines the importance of overcoming 'standardised' or 'normalised' ideals. And instead to allow for different and alternative ways of being. Accordingly, indeterminacy can also be an instrumental value, such as promoting an open and pluralistic society. But I leave that open for further research.

1.4 Conclusion

In order to show how the provided self is understood, I first discussed Aristotle's essentialism and Descartes' mind-body dualism. Both conceive of the self as static, invariable and independent. Moreover, Descartes' distinction between inside and outside reinforces the idea of an independent and autonomous self. In doing so, it sustains the impression that technology has no effect on the self, and vice versa. Further, this understanding excludes the possibility of becoming and (radical) change.

To overcome these shortcomings, the interactionist self with focus on Nietzsche's will to power ontology was presented. For Nietzsche, the self has no essence, but is inherently indeterminate. This means, it is a priori undefined and its development also remains unfinished and unknown. The self thus never 'is', but is always in a state of becoming formed through

interactions.

In this, the self is understood as a will to power organisation. It is an organisation (like a system) and the result of organisation (as activity). Since power is only power in relation to other power, the self is a will to power organisation among (but not separate from) other will to power organisations. Given this dynamic, there is a constant chaos at work. Chaos or tension is required for the will to power organisation in order to grow and prevent decay. At the same time, however, the will to power organisation must be able to organise that tension including the power struggles with others. Otherwise the self cannot form a seemingly stable and independent unity. The challenge, then, is to maintain the discrepancy between intense tension or chaos and good organisation.

The relational and dynamic nature of the self implies that it is always constrained to some degree in the process of becoming by the context in which is embedded. And given the prevalence of technology that transport ideals and values, it is part of self-formation. Or rather, technology becomes part of the self. Thereby, technology neither has no influence on the self nor does it control the self completely.

Ultimately, self-formation is understood as a conscious and intentional process that enables the self to reflect on and identify with certain ideals or assigned identities and negotiate their meaning accordingly. Especially in view of the indeterminacy of the self, constant re-negotiation is necessary because the self can never be fully defined. This also means that ongoing struggle or negotiation is required for the self to grow. Accordingly, the indeterminacy of the self is desirable in that it allows for different ways of being rather than conforming to 'standardised' ideals. The assumption is that this promotes the well-being of the self.

In the following chapter 2, I will elaborate on the language model called Generative Pre-trained Transformer 3, in short GPT-3. In doing so, I will emphasise which ideals the synthetic texts convey. In chapter 4, I will then examine how GPT-3 challenges the conditions for self-formation.

Chapter 2

The Workings and Worldview of GPT-3

Self-formation increasingly becomes technological self-formation. Thus, in this chapter I will present the large language model called Generative Pre-Trained Transformer 3 (GPT-3). It was released in 2020 and is capable of writing articles, summarising texts and producing computer code.

To better contextualise GPT-3 and explain the underlying functioning of the model, I will first provide a brief history of *machine learning*. That is, how models have evolved from hard-coded rules to finding patterns in data. Based on that I will turn to the details of how GPT-3 'learns' a language. This is important because regardless of how astonishing the output may be, we must remind ourselves that GPT-3 relies on the mathematical formalisation of language. As I will also show, this seemingly objective endeavour leads to value-laden results, such as texts that associate women with less power. GPT-3 ultimately contains a particular and, in a sense, 'normalised' or 'standardised' worldview, which it transports across contexts. This is important because language conveys ideals, beliefs and power structures (chapter 3).

2.1 What is GPT-3?

The Generative Pre-trained Transformer (GPT) is a family of language models developed by the company OpenAI. The first version, GPT-1, was introduced in 2018 (Radford, 2018; Radford et al., 2018), a year later GPT-2 was released (Radford, Wu, Amodei et al., 2019; Radford, Wu, Child et al., 2019) and in 2020 its successor GPT-3 appeared (Brown et al., 2020).[1]

The main differences between the models are the size of the training data used and the amount of parameters the model has. The size of the model is insofar important as it is linked to the quality of the output. Generally speaking, the larger the model, the 'better' the result (Brown et al., 2020, p. 4).[2] Accordingly, GPT-3 produces higher quality texts compared to GPT-1 and GPT-2.

So while GPT-1 had 117 million parameters, the number increased to 1.5 billion in GPT-2, and again increased to 175 billion parameters in GPT-3.[3] Similarly, the training data increased from 5GB of text (GPT-1), to 40GB (GPT-2), to 570GB (GPT-3). This makes GPT-3 one of the biggest language models to date. In a test study with 62 participants, for example, the mean accuracy in correctly attributing the author of news articles of about 500 words was 52%, which is equivalent to guessing (Brown et al., 2020, pp. 25–26). In other words, in some cases it is difficult to distinguish whether the text was written by a GPT-3 or a human.

To generate a text, a task description is provided to GPT-3, which is

[1] In January 2022 OpenAI relesead a refined version of GPT-3, called 'InstructGPT' (Lowe & Leike, 2022).

[2] At least that is the assumption that larger language models are better, as the trend towards bigger models also shows. This does, however, not mean that a small model will necessarily perform poorly on a particular task.

[3] GPT-3 itself is also a set or family of models. This means that there are different models of different sizes (Brown et al., 2020, p. 8). Usually, 'GPT-3' refers to the model with 175 billion parameters.

then processed. The Guardian, for example, published an article that was completed by GPT-3 with the following instruction:

> Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI. (GPT-3, 2020)

Simply put, GPT-3 works like a large auto-complete function. In section 2.3.1, I will elaborate on the underlying *transformer architecture* (Vaswani et al., 2017) through which GPT-3 learns a linguistic representation and generates text accordingly.

Next to generating synthetic texts, such as news articles or poems, texts can be summarised or translated, or questions can be answered. It is also possible to create games or computer code (Dale, 2021, p. 115). And in combination with other models, GPT-3 can produce paintings and images (e.g., Patashnik et al., 2021; Ramesh et al., 2021; Reed et al., 2016).

In general, GPT-3 can be used to perform various natural language processing (NLP) tasks. GPT-3 does this without being specifically trained for certain tasks (see section 2.3.2). Against this backdrop and in combination with some (impressive) sample texts, GPT-3 attracted a lot of media attention. Speculations ranged from the claim that GPT-3 would change the field of machine learning to the assertion that it would produce nothing useful because it lacks understanding of meaning (Bender & Koller, 2020; Marcus & Davis, 2020).

2.2 A Brief History of Machine Learning

In order to understand how we arrived at sophisticated models like GPT-3, I will briefly sketch how the field of machine learning started and how it has developed in recent years. Although today's models are able to achieve things that were unimaginable a few years ago, such as synthetic texts by GPT-3, we must not forget that they still work through mathematical operations and the

calculation of probabilities. These probabilities ultimately determine what GPT-3 generates.

2.2.1 Imitation of Intelligence

In 1956, John McCarty and Marvin Minsky hosted the Dartmouth conference, which is considered as the birth-place of the research field of *Artificial Intelligence* (AI) – the term coined by John McCarthy (Nilsson, 2009, pp. 52–56). Researchers from various fields, such as cognitive science, psychology, neuroscience and computer science took part in this summer project. Their shared interest was to understand how human behaviour and cognition functions and whether it could be taught to a machine. Overall, their vision was lead by

> the conjecture that every aspect of learning or any other feature of intelligence can be in principle so precisely described that a machine can be made to simulate it. (McCarthy et al., 2006, p. 12)

The term Artificial Intelligence is controversial and there is no single definition as it refers to various sub-disciplines, ranging from image recognition, language translation, robotics, or even industrial mechanisation (Nida-Rümelin, 2022, p. 71).[4] Regardless, the overarching goal of these different branches is to quantify, formalise and represent human knowledge and behaviour so that a machine can mimic it and perform a specific task (Kowalski, 1979). Accordingly, the Webster's Dictionary describes AI as 'the capability of a machine to imitate intelligent human behaviour'[5] (e.g., Turing, 1950).

[4]Crawford (2021, p. 7) even claims that Artificial Intelligence is 'neither artificial nor intelligent'.

[5]See https://www.merriam-webster.com/dictionary/artificial%20intelligence

2.2.2 From Rules

For a long time the game of chess was considered as the guiding principle to build intelligent machines. Partly, because being able to play chess was considered as intelligent behaviour. But rather because chess is neither too simple nor too complex to formalise mathematically (Ensmenger, 2011, p. 6). The chessboard consists of a fixed number of possible places, and follows clear rules. This makes it possible to relate cause and effect, from which moves can be derived and calculated with increasing accuracy through more data. In this sense, chess offers a bounded problem space with ideal conditions.

The idea that intelligence comes down to merely following rules was also further developed by Herbert Simon and Allen Newell. For them, the human mind and the machine were both 'complex information processing systems' (Dick, 2015, p. 625). Although this similarity was not meant as a clear comparison, they believed that human behaviour can be reduced to a set of formal rules and that these rules can be taught to a machine (Dick, 2015, p. 626).

The formalisation and mathematical representation of knowledge was also the leading principle in so-called *expert systems* that researchers began to build in the 1970s (Winograd, 1990, p. 170). These systems contain a specific 'knowledge base' from domain experts, e.g. in the medical sciences (Forsythe, 1993, p. 451). Based on rules in the form of *if...then*, the expert system is able to process an input (i.e., a problem) and produce an output (i.e., a decision).

An early example in the field of natural language processing is the programme named ELIZA, developed by Joseph Weizenbaum in 1966.[6] ELIZA is one of the first chat bots that is able to respond to text messages from a human. But instead of understanding the meaning of the conversation, which would allow to respond in a situational and contextual manner, ELIZA

[6]The original source code of ELIZA was published under https://sites.google.com/view/elizagen-org/the-original-eliza

simply followed rules or 'scripts'. A script determines how ELIZA responds based on 'keywords'. Often these are simple follow-up questions (e.g., why do you think so?) to statements made by the individual (Weizenbaum, 1966, p. 37).

As this early example already illustrates, it is apparently sufficient to formalise language and calculate the likelihood of a subsequent word or phrase by merely matching keywords or word patterns.

2.2.3 To Patterns

The approach to train the machine certain rules that are based on human knowledge is known as *symbolic AI*. In contrast, *subsymbolic AI* that emerged around the same time, does not rely on explicitly defined rules. Instead, it learns these rules or statistical correlations itself by finding patterns in the data.

Due to the increase of available data and the increase of computational resources, a shift from symbolic AI to subsymbolic AI took place. This is also what has become known as machine learning. A consequence of this shift is that the initial interest of the Dartmouth conference was to understand human cognition, but:

> By the late 1960s, most researchers in pattern recognition ultimately cared little whether neural networks in any way replicated human cognition; the networks were tools for prediction, not means for understanding the brain. (Jones, 2018, p. 677)

In order for the machine to find patterns (i.e., to *learn*), the system is presented with data in the so-called training phase (see C. M. Bishop, 2006, p. 2). Training can thereby happen in a supervised or unsupervised manner.[7]

For supervised learning, developers present the machine learning system data with labels, which represent the desired output. Next to that, certain

[7]Further learning methods include reinforcement learning and semi-supervised learning.

characteristics or features need to be defined that adequately represent the phenomenon. This is also known as feature extraction. So, for example, if we present a sufficient amount of emails with appropriate labels such as 'spam' and 'non-spam', including the features (i.e., list of words) that represent each class, the algorithm will learn to discern between these two categories.

It does so by learning the relationship between a set of input data $X = (x_1, ..., x_n)$ and a set of desired output or classes $Y = (y_1, ..., y_n)$ by approximating a target function $f : X \rightarrow Y$. In its simplest form the function looks like $Y = f(X)$. This allows to calculate the probability of Y given X: $P(Y|X)$.

So instead of defining explicit rules like *if this word appears, then the email is spam*, it is possible to define certain words (i.e., features) that represent spam or non-spam. The machine learning model then learns the correlations accordingly, such as $spam = f(< training\ data >)$. Through several iterations in the training phase, the so-called weights, or parameters, of the function are adjusted in order to reduce the error rate (T. M. Mitchell, 1997, pp. 10, 88). In that sense, the adjustment of the parameters represent the learning (i.e., fitting) of the model through 'experience'. Later, during the testing phase where the model is validated these parameters can be adjusted further.

Once the target function performs well, it can discern or classify further examples it has not yet seen by calculating a probability: $spam = f_{model}(< unseen\ email >)$ or $P(spam|email)$. If a certain threshold is reached, the model can classify the email into one of the available labels. Thereby it is important that the model is able to generalise from the data. If it would simply learn the training data one by one, *overfitting* occurs and the model would not be able to adequately classify emails containing words it has not seen. On the contrary, if there is too little data in the training set, then the problem of *underfitting* occurs, in which case the model does not learn a sufficient representation of how input and output are related (Frické, 2015, p. 654).

For unsupervised learning, on the other hand, the data is unlabelled. The data is thus unstructured and the learning approach is more exploratory. That is, the model groups or clusters data based on similar characteristics. For the email example, an unsupervised model would not be able to discern between spam and non-spam (as these labels are not present). Instead it could group emails based on other characteristics, such as sentiment. I will return to unsupervised learning, as GPT-3 is based on this approach.

Deep Learning

Around 2011, deep learning models, a subset of machine learning, took shape, again due to the increase in computing resources (M. Mitchell, 2020, p. 2). In their basic form, deep learning models are neural networks consisting of an input layer, an output layer and a variable number of hidden layers. Here, *deep* refers to the amount of hidden layers, i.e., the depth of the network (Goodfellow et al., 2016, pp. 168–9). The units of one layer are connected to the units of the following layer. Where the connections between the units are the parameters or weights. For GPT-3, this is 96 layers and 175 billion parameters (Brown et al., 2020). To put this into perspective, a neural network that consists of two units in the first hidden layer, three units in the second hidden layer, and one output layer, would result in a 3 layer network with 9 parameters.[8]

Deep learning models differ from earlier models in that they can select (high-dimensional) features themselves. To do so, they identify patterns in the data during the training phase. The more data the deep learning model processes, the better it can optimise the target function through the adjustment of its parameters, that is the strength of the connection between different units. In case of the previous example of discerning emails into spam and non-spam based on a list of words (i.e., features), the definition of features is no longer relevant. Instead, it is sufficient to present the model

[8]The input layer is not counted.

with a large amount of labelled emails from which it can extract features and discern between the two classes on that basis.

An implication of the subsymbolic approach, however, is that the model might discern emails based on other features, as humans would. In other words, the features or patterns that the model picks up are most likely different from how humans would represent the phenomenon. Simply because the machine 'sees' information as a large set of numerical values (see section 2.3.1). A vocabulary consisting of the words *queen* and *king*, for example, could be represented as follows:

$$queen = \begin{bmatrix} 1 \\ 0 \end{bmatrix} king = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

While this is a simplified example, the high-dimensional features are often inconceivable and incomprehensible to us humans. In the end, there is a lack of shared understanding. I will return to this in section 4.2.2.

2.3 The Functioning of GPT-3

Compared to other deep learning models, GPT-3 is special in at least two ways. First, because of its generative nature. And second, because it is task-agnostic, meaning that it can perform a variety of tasks without being specifically trained to do so. The first aspect is not remarkable in itself, as GPT-3 is not the only generative model.[9] But in combination with the second feature, the strength of the language model unfolds.

[9]See for example https://thispersondoesnotexist.com/. This project uses a so-called generative adversarial network (GAN) to create images of faces that do not exist. Another example is David Cope's model, which composes music in the style of Bach.

2.3.1 Generative Model

In general, there are two types of machine learning models, namely discriminative and generative (Ng & Jordan, 2001). As the name Generative Pre-trained Transformer suggests, GPT-3 is a generative model. But so far, I have discussed discriminative models. Simply because they are slightly easier to grasp.

A discriminitave model calculates on the basis of a labelled data set, the probability of class y of a data point x by the conditional probability: $P(y|x)$. Generative models, on the other hand, learn to calculate both, the probability of the data point x and of the class y, that is the joint probability $P(x, y)$. To do this, they create their own representation of the training data by learning relevant features. Based on similarities the data is grouped or categorised into classes $P(y)$. This is especially beneficial for unsupervised learning tasks, such as language modelling, where classes are not available (i.e., unlabelled data) (Radford et al., 2018, p. 1). Based on the learned representation, the model can predict the probability of the data point x: $P(x|y)$.

For the email example, this means: The discriminative model directly calculates the probability that an email is spam (y) given the words in the email (x): $P(spam|email)$. It does so based on the learned *boundaries* between the two classes it derived from the labelled data. The generative model, in contrast, learns the features (x) and based on that models the distribution of classes y. In case the labels are present, it can compare the assumed classes with the actual class y. In case the labels are not present, it can group emails based on similarity. In the end, the model can then also generate new synthetic emails (x') that are similar to the training data (x).

So while both discriminative and generative models can be used to categorise data, only the generative model can create synthetic data. This is because discriminative models learn the decision boundaries between classes, whereas generative models learn the distribution of classes (see Figure 2.1). In view of this, discriminative tasks are often computationally cheaper as

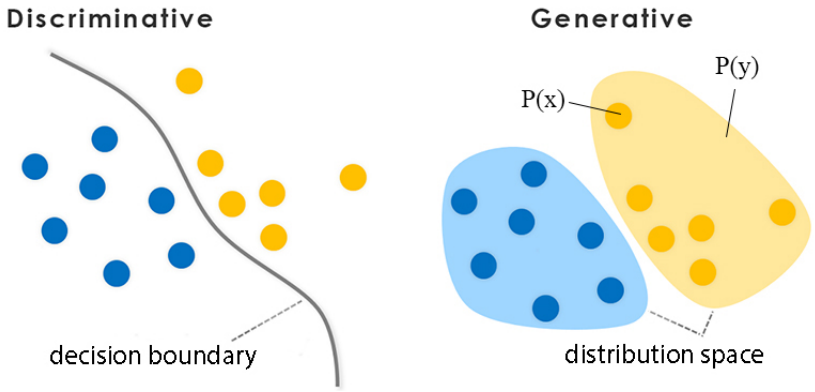

Figure 2.1: Discriminative and Generative models. Discriminate models learn the decision boundaries and the conditional probability $P(y|x)$. Whereas generative models learn the distribution space through the joint probability $P(x, y)$ (adapted from PRIMO.ai, 2020).

they require less data (Radford et al., 2018, p. 1).

Word Embeddings

For a model to create a linguistic representation, so-called word embeddings are used. In this, each word is represented as a vector, which contains the coordinates of a point in a (multi-dimensional) space. This space is defined by the training data. And the more similar two vectors are, the closer the words are to each other.

Word2Vec was, and still is, a common method to learn these associations (Mikolov, Chen et al., 2013). For example, in vocabulary that consists of the words 'queen', 'king', 'female' and 'male', each word is converted into a word vector, also representing the relationship between the words (M. Mitchell, 2020, p. 238).

$$queen = [1, 0, 0, 0],\ king = [0, 1, 0, 0]$$

$$female = [0, 0, 1, 0],\ male = [0, 0, 0, 1]$$

This would then allow for a mathematical operation like $king - man + woman = queen$ (Mikolov, Yih et al., 2013, p. 749). This is of course a

simplified example, but

> The idea here is that once all the words in the vocabulary are properly placed in the semantic space, the *meaning* of a word can be represented by its location in this space – that is, by the coordinates defining its word vector. (M. Mitchell, 2020, p. 241)

The assumption is that the meaning of a word is derived from the words that surround it. A common quote in this context is: 'you shall know a word by the company it keeps' (Firth 1957, cited in Hutchinson et al., 2020, p. 5494). This means, the data distribution or semantic space and the vectors it contains are responsible for what LLMs can generate. The larger the model, the larger the semantic space, the better the representational accuracy of the relationship between the words (Mikolov, Chen et al., 2013, p. 10).

Transformer Architecture

Generally speaking, GPT-3 relies on the same principle, namely the numerical representation of words or rather sequences of words. To do so, it uses the *transformer architecture* (Vaswani et al., 2017).

In contrast to word2vec, the transformer processes sequences of words instead of considering word by word in isolation. This is relevant in that a single word can have different meanings depending on where it stands in the sentence. Very simplified, for example, the word 'good' does not automatically imply something positive, since something could also be 'not good'. Further, a single word can also have multiple meanings, such as the words *bat* (animal / racket), or *pupil* (student / eye) and so on. So compared to previous methods for learning word embeddings, transformer networks 'capture higher-level semantics' (Radford et al., 2018, p. 2).

To do so, the transformer first takes a provided input sequence $(x_1, ..., x_n)$ and creates a compressed representation or embedding $(z_1, ..., z_n)$ based on the pre-trained 'base knowledge'. This is done by the so-called *encoder*. Importantly, the position or context of each word within the sequence in relation

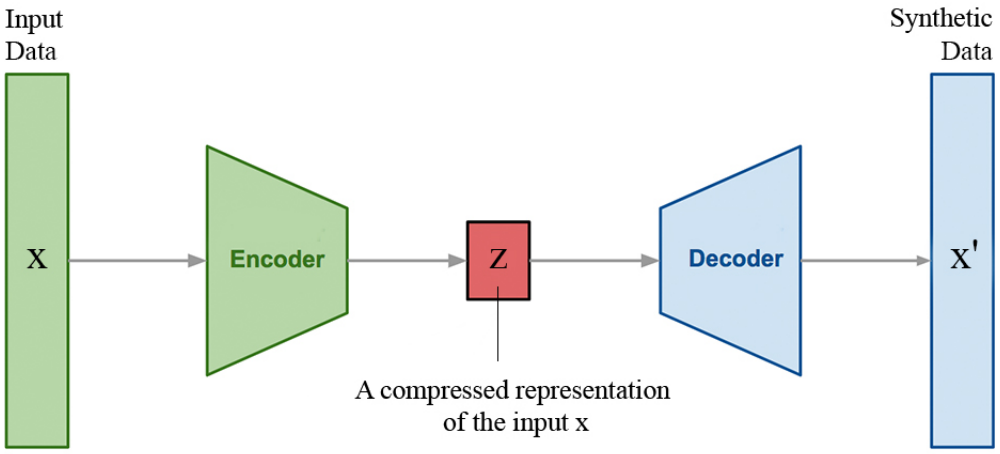

Figure 2.2: A simplified encoder-decoder architecture on which the transformer is based on (adapted from Weng, 2018). For a more detailed illustration of the transformer architecture see Vaswani et al. (2017, p. 3).

to the preceding and succeeding words is also encoded. This is known as *attention mechanism*. In the sentence 'She is eating a green apple', for example, the word pair $[eating, apple]$ has a higher attention than $[eating, green]$. The *decoder* then takes that representation z and generates an output $(x'_1, ..., x'_n)$ accordingly (Vaswani et al., 2017, p. 2) (see Figure 2.2).

2.3.2 Task-Agnostic Performance

Usually, machine learning models are limited to a very specific use case, such as playing chess, recognising melanoma skin cancer in medical images, or classifying loan applicants. This is especially true for supervised learning models, as the model is optimised for a specific purpose.

GPT-3, however, is a so-called *few-shot learner* with a task-agnostic architecture (Brown et al., 2020). This means GPT-3 has been 'pre-trained'[10] in an unsupervised manner with a large amount of unlabelled data. From

[10]Hence the name Generative Pre-Trained Transformer.

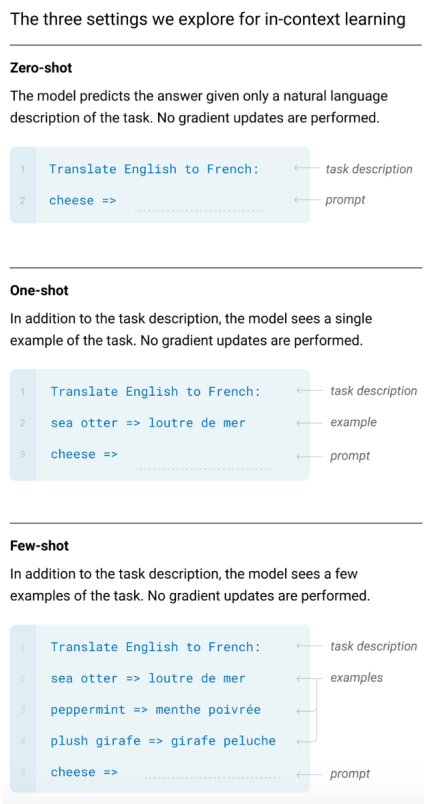

Figure 2.3: Zero-shot, one-shot and few-shot task settings (Source: Brown et al., 2020, p. 7).

this data, GPT-3 then extracts relevant features and learns statistical correlations of word occurrences by itself. This 'base knowledge' or representation, so to speak, allows GPT-3 to solve other tasks for which it has not been specifically trained (i.e., downstream tasks). As Brown et al. (2020) put it:

> During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. (p. 3)

To solve a task, we can provide an example of how to solve the task in addition to the task description. Depending on how many examples are provided, the method is called few-shot, one-shot, or zero-shot (see Figure 2.3).

As Brown et al. (2020) explain:

> The main advantages of few-shot are a major reduction in the need for task-specific data and reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset. The main disadvantage is that results from this method have so far been much worse than state-of-the-art fine-tuned models. (p. 6)

A fourth method would be to fine-tune GPT-3 for specific tasks. For this purpose, the parameters can be adjusted using labelled data. So GPT-3 is trained in a supervised manner in addition to the previously unsupervised pre-training, making it semi-supervised. A major limitation, however, is that each task would require a large data set (Brown et al., 2020, p. 6). Regardless, GPT-3 itself is not fine-tuned, although this is possible in principle.

In general, the task-agnostic architecture becomes apparent considering the various natural language processing tasks GPT-3 is able to perform. Such as, summarising or translating texts, writing news articles or poems or answering questions (e.g., chat bot). Besides, Microsoft, which has an exclusive license for GPT-3 (Scott, 2020), recently introduced its first products that use the language model to 'help users build apps without needing to know how to write computer code or formulas' (Langston, 2021).

But despite this wide use range and the (rather) astonishing results, the generative capabilities of GPT-3 are still limited to the learned distribution space. Hence, the data that GPT-3 is pre-trained on becomes increasingly important.

2.4 The Importance of Data Sets

Although deep learning systems learn statistical correlations by themselves, they still rely on intensive human labour, ranging from the construction of the material infrastructure, from collecting and preparing the data set, to

constructing and validating the model itself (Crawford, 2021; Crawford & Joler, 2018). Especially during model construction and validation, the goal of the machine learning model is set by engineers. They make judgements about which data to use for training and what parameters and thresholds to adjust to define performance.

It is primarily data that sets the epistemic boundaries of current deep learning models (Crawford, 2021, p. 98). In other words, data determines how the model 'sees' the world and thus what it is capable of doing. For example, as Kate Crawford (2021, p. 97) mentions, a system that is trained to discern between apples and oranges, but the data set only contains green apples, it is not able to recognise a red apple, at least not adequately. This is because the system assumes 'all apples are green'. In this case underfitting occurs where no (adequate) link between input and output can be established. Similarly, the data that GPT-3 is trained on is responsible for how the input embeddings are created and what it generates accordingly.

2.4.1 The 'Normalised' Self

The amount of data (570GB) that GPT-3 was trained on shows the underlying assumption that more data leads to 'better' results (boyd danah & Crawford, 2012, p. 663). And in general, models can optimise the target function that links input and output by processing more data (see T. M. Mitchell, 1997, p. 2). But despite the large amount, this does not mean that the sequences of words contained and the viewpoints associated with them are diverse (see Bender et al., 2021, p. 613).

Scientists and engineers always operate within a certain 'disciplinary perspective' (Boon, 2020). This means certain background assumptions and tacit knowledge shape decisions in both the model construction and validation phases. This is crucial, because data used to train deep learning systems often represent the viewpoints of the engineers (Raji et al., 2021, p. 8). And considering that current models are mostly developed by white, western men, machine learning models embed a very specific set of values (Cave & Dihal,

2020; Crawford & Paglen, 2019; Denton et al., 2021).[11]

For GPT-3, the largest part of the training data (60%) comes from the web scraping repository Common Crawl[12] (Brown et al., 2020, p. 9). The exact sources of the Common Crawl data set are not disclosed, and given the sheer amount it is difficult to analyse the repository appropriately. But initial analysis indicate that Common Crawl 'contains a significant amount of undesirable content, including hate speech and sexually explicit content' (Luccioni and Viviano, 2021). Another large part (22%) comes from WebText2 data set, which was also used to train GPT-2. WebText2 collects data from Reddit, where demographics are primarily associated with white young men (Bender et al., 2021, p. 613). This leads to certain views and assumptions being over-represented, known as representation or sampling bias (Blodgett et al., 2020, p. 5455).

Representation bias may occur unintentionally because of the disciplinary background assumptions. But engineers often overlook that these models are not objective or 'universal truths' (Forsythe, 1993, p. 449). Instead, they primarily focus on technical aspects such as performance, generalisation or efficiency (Birhane et al., 2021, p. 3). As Birhane et al. (2021) further elaborates:

> The upshot is that the discipline of ML is not value-neutral. We find that it is socially and politically loaded, frequently neglecting societal needs and harms, while prioritizing and promoting the concentration of power in the hands of already powerful actors. (p. 10)

In view of this, several studies suggest that GPT-3 contains and generates various harmful and stereotypical beliefs.

---

[11] In psychology research, too, the representatives from whom human behaviour is derived are primarily White, Educated, Industrialised, Rich and Democratic (WEIRD) (Heinrich et al., 2010).

[12] https://commoncrawl.org/the-data/

For example, GPT-3 creates texts that contain gender stereotypes. 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). It also assumes that women have professions such as nurse, receptionist or housekeeper, and men pursue professions associated with a higher levels of education, such as banker or professor (Brown et al., 2020, p. 36). Further, Brown et al. (2020, p. 37) note that the 'Black' race shows consistently low sentiment. 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 with the mentions of disability, such as linking mental illness with homelessness (Hutchinson et al., 2020). In summary, GPT-3 conveys particular (and harmful) beliefs and assumptions in relation to gender, race, religion and ableism.

While data is always a partial representation of the phenomenon, here natural language, representing a phenomenon is nevertheless a highly normative task.[13] This is because it takes 'decisions about what entities to include in and exclude from a representation' (Harvard and Winsberg, forthcoming, p. 3). For example, as boyd danah and Crawford (2012) mention, Twitter is a common source for scraping text, but:

> Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous. (p. 669)

Representation is thus an act of power, for it is a matter of judging what is considered desirable, adequate or 'normal' (e.g., Ames, 2018, pp. 2–3;

[13]In a broader view, not only is the contextual or domain-specific type of language important, but the representation of different languages itself is also important. '[O]ver 90% of the world's languages used by more than a billion people currently have little to no support in terms of language technology' (Bender et al., 2021, p. 612).

Gururangan et al., 2022; Raji et al., 2021, p. 8; West, 2020). In doing so, normalisation ultimately excludes everything that does not fit into the defined categories and into numerical values at all.

In terms of self-formation, this is crucial because GPT-3 charges and confronts the self with ideals and values. Thereby, the reduced and stereotypical understanding of the self can lead to psychological harms through discrimination and exclusion (Weidinger et al., 2021, pp. 13, 15). And on the basis of certain assumptions, opportunities could be unfairly denied to the self, known as allocational harms (Blodgett et al., 2020, p. 5455). Thereby, those most vulnerable to the ideals embedded in GPT-3 are already underprivileged or marginalised within society (e.g., Bender et al., 2021; Benjamin, 2019; Birhane, 2021; Eubanks, 2019; O'Neil, 2016). This, in turn, maintains current power structures, i.e., the status quo.

Ultimately, language categorises and thus structures reality (chapter 3). This is also true for GPT-3. In doing so, given its widespread use across contexts, GPT-3 conveys and perpetuates a particular and, in a sense, 'normalised' or 'standardised' view of the self. This does not mean that this is the 'true' self, but as I will show in chapter 4, it is increasingly difficult for the self to challenge and negotiate these transported meanings.

2.5 Conclusion

To better contextualise the workings of the language model GPT-3, I provided a brief history of machine learning. Scientists and engineers perceived of 'intelligence' as following rules. Accordingly, they built rule-based systems that relied on human knowledge. Today, deep learning models learn correlations themselves by analysing large amounts of data.

For GPT-3 to learn a linguistic representation, it encodes word sequences in the form of words embeddings. In other words, it correlates word occurrences. Based on this representation, GPT-3 is able to generate synthetic data. Other tasks include summarising or translating texts, or answering

questions. Despite the (rather) impressive output and the wide range of tasks it can perform, the quality of the training data is particularly important as it sets the epistemic boundaries of GPT-3.

Given the mathematisation of language and the large corpus of text that was used for training, this seems to be an objective and neutral undertaking. But engineers still have to make decisions about what data to use and how to validate the model. This introduces values trough tacit background assumptions. Further, lots of data does not imply diversity of the data. Since GPT-3 was mainly trained on data scraped from the Internet, it holds a very specific representation of language that mainly refers to young male internet users. In the end, GPT-3 is not value-free, but creates texts with social stereotypes and prejudices related to gender, race, religion and ableism.

This is problematic because the widespread use of GPT-3 transports these particular ideals across contexts. In a sense, GPT-3 reduces and

normalises the self. And in doing so, it excludes alternatives that do not fit into numerical representations.

Before I discuss the challenges that GPT-3 poses to self-formation in chapter 4, I will first turn to the role of language. As I will show, language is an interactive task, that is contextual and ambiguous. Since language is a prerequisite for interaction, the self and language are intertwined and co-constitute each other.

Chapter 3

Language and Self-Formation

So far I have shown that the self is formed by the interactions it has with others. As the self increasingly interacts with LLMs, it is likely that they have a great influence on self-formation. Especially since, GPT-3 generates synthetic texts and thus language.[1]

We use language, whether written, spoken or sign language, to communicate and share ideals and beliefs. This means language enables interaction. Moreover, metaphors like 'I need to blow of some steam', or 'I need to recharge my batteries' illustrate that technologies shape language. In doing so, these metaphors also shape how the self experiences itself. In this sense, technology, language and the self are intertwined. And while previous technology mediated understanding and thus (passively) shaped language, GPT-3 not only mediates but increasingly (and more actively) constitutes language (see Burrows, 2009, p. 451).[2] For this reason I think the theoretical framework of the interactionist self can benefit from considering language. To address

[1]Next to synthetic texts, GPT-3 also generates computer code, which plays an increasingly important role in our technological environment.

[2]This is not to say that GPT-3 produces a new language, although some output may well be astonishing, whether by 'error or genius' (Elkins and Chun, 2020, p. 5). Rather, its generative capabilities distinguish it from earlier technologies.

the question of *what role does language play in self-formation*, I will turn to Ludwig Wittgenstein and the notions of *language games* and *forms of life*.

Since the implications of the co-constitution of language and GPT-3 will only become clear in the coming years (if ever?), this is undoubtedly beyond the scope of this thesis (e.g., Mooney & Evans, 2015). The following discussion on language therefore provides a preliminary conceptual tool that contributes to the analysis of this co-constitution. In doing so, I want to highlight the relevance of the triadic interaction between the self, language and GPT-3.

3.1 Language Games and Forms of Life

With the concepts of *language games* and *forms of life* Ludwig Wittgenstein emphasised the importance of language use.

The language we use is like a game we play. That is, it follows 'rules' in the form of conventions and customs, and the self changes its actions and behaviour accordingly. For example, a different language game is played when using the word 'water' in a restaurant than when confronted with a small fire. The former could be a (polite) answer to a question, the latter is more of a (nervous?) command. Thus, one and the same utterance can lead to different actions based on the situation. Importantly, this also means that language and its meaning cannot be reduced to words themselves (Coeckelbergh & Funk, 2018, p. 168). Rather, language is interwoven with contextual and situated use. Wittgenstein thus notes:

> I shall also call the whole, consisting of language and the activities into which it is woven, a "language-game". (cited in Coeckelbergh, 2018, p. 1505)

A language game is always performed in a larger context. This is what Wittgenstein calls a *form of life*. A form of life can be society, different (professional) disciplines, or different social relations in general. A form of

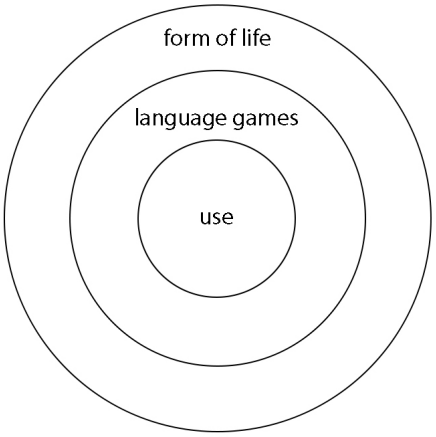

Figure 3.1: Language use, language games and form of life (adapted from Coeckelbergh, 2018, p. 1506).

life is thus understood as a social and cultural context with its own customs, values and assumptions (Coeckelbergh, 2018, p. 1506). Ultimately, language use is structured by language games, which in turn are structured by a form of life (see Figure 3.1).

In every form of life, different language games (e.g., activities or interactions) are performed. Still, different forms of life can structure the same language games and thus language use in different ways. For example, if we consider greeting as a language game, then this game is probably played differently with a close friend than with a casual acquaintance. Accordingly, the self also acts differently. In other words, language shapes the interaction, while interaction shapes language. This is because the rules of a language game prescribe a certain behaviour, how activities should be carried out (Coeckelbergh, 2018, p. 1508). In that sense, the will to power organisation (i.e., the self) organises itself according to the rules of the game being played.

The form of life, however, is not external to the use of language, nor does it cause the use of language (Coeckelbergh, 2018, pp. 1506, 1514). For the use of language also constitutes a form of life. For example, describing the self as a 'citizen', 'user' or 'customer' in a particular situation involves and conveys different forms of life, each of which entails different expectations,

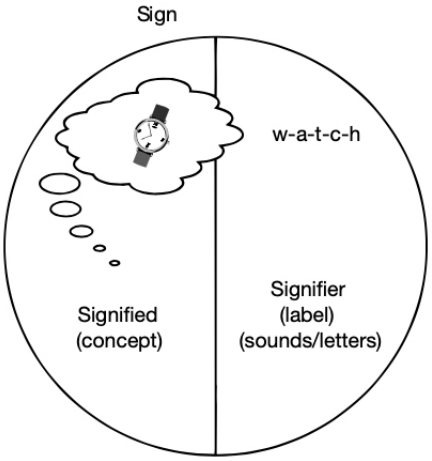

Figure 3.2: Saussure's model of the sign (Source: Mooney & Evans, 2015, p. 21).

beliefs and assumptions (see Mooney & Evans, 2015, p. 32). Wittgenstein thus states, 'to imagine a language means to imagine a form of life' (cited in Coeckelbergh & Funk, 2018, p. 169). The use of language, language games and form of life ultimately constitute each other.

3.1.1 Public Language

Due to contextual and situated use, language is ambiguous. For example, the word 'light' can refer to physical weight, the intensity of colour, or to a source of illumination. In each case, the word functions as a *signifier* representing a concept, i.e., the *signified* (see Figure 3.2). In order to reduce ambiguity (of the signified) and make interaction meaningful, a shared understanding is necessary, which in turn presupposes a form of life (Coeckelbergh & Funk, 2018, p. 171).

The self is always embedded in a form of life, in which language already exists. Even if the self were to develop its own language with its own rules, the self would have to change existing symbols to do so. This refers to Wittgenstein's argument that there is no private language. In other words, language is ultimately social and public.

As a consequence, many concepts (i.e., signified) are based on tacit knowledge (Coeckelbergh & Funk, 2018, p. 172). This is because many rules (beyond grammar) are learned implicitly rather than explicitly by listening to others and using language (i.e., performing language games). In doing so, we develop a practical know-how; similar to observing a board-game long enough (Coeckelbergh & Funk, 2018, p. 169). This observation, however, is not a passive endeavour, rather it relies on interaction and joint attention (Bender & Koller, 2020, p. 5190).

Moreover, as Mark Johnson (2008, p. 162) points out, we often use our body as a frame of reference to make sense of time and space. For example, we say someone hides 'behind' the door or someone puts the volume 'up'. Thereby, these spatial metaphors vary in different forms of life. As such, the future can be 'in front' of us, 'west' of us, or 'uphill' (Cooperrider & Núñez, 2016). Next to the tacit knowledge, this example also illustrates that language is an embodied endeavour.

Given this social and embodied nature of language, and contrary to Descartes's understanding, the 'inside' and the 'outside', the mind and the body, the self and the other, are not separate, but are in a dynamic and reciprocal relation (see Husserl, 1970, p. 82). As a result:

> Cognition and thinking do not belong to an ontologically separated substance, as Descartes at least suggested with his argument in the Meditations; instead, in performance, body and mind are together. This also means that language and thinking are innately linked to the body, are deeply embodied, at the level of performance. (Coeckelbergh & Funk, 2018, p. 175)

As such, language can influence object recognition (Boutonnet et al., 2013) or colour perception (Winawer et al., 2007). This means, language cannot be de-coupled from use, so that language constitutes thought and action of the self.

Accordingly, the language we use to describe something influences how we make sense of that thing or phenomenon. For example, the computational

metaphors 'the brain is a computer' and 'the computer is a brain' are widely used in both research (Baria & Cross, 2021) and public discourse. If the self sees its brain as a computer, it might understand itself as processing and storing information and assume that it needs to process more and more information, or it could conceive of its body parts as hardware that can be replaced. Language ultimately shapes the way the self experiences and understands itself, which in turn affects self-formation. Hence, synthetic texts generated by GPT-3 also shapes how the self views itself.

3.1.2 Social Categories

Through shared understanding, language allows to form communities (e.g., Alim, 2006, pp. 72–3). That is, the self can relate to others and establish a sense of belonging (Firth, 1971, p. 42; Zaidel, 2018, p. 27). But language also allows the self to distance itself from others. In other words, the self identifies itself in relation to others, either through similarities or through differences. This means, language is categorical (e.g., Blake, 2016) and functions as the signifier of ideals and values (Coeckelbergh, 2018, p. 1508). This is relevant insofar as:

> Identity comprises group membership and self-definition in terms of social categories, including nationhood, community, sense of place, and ethnic and religious identity, where these are salient. It defines who shall be deemed ingroup and outgroup, and therefore, what shall be the basis for sharing symbols and metaphors with others. It also includes self-identity, in which adherence to particular values or beliefs becomes part of the self: "I am the kind of person who believes such and such." (Haste, 2004, p. 420)

Again, through language, the self understands itself (in relation to others and its environment). This understanding, however, is not fixed or static. Rather, the social, contextual and ambiguous character of language allow for interaction and negotiation (i.e., power struggles), through which the self can

form itself. LLMs, on the other hand, convey certain ideals from a specific context (i.e., the Internet) across contexts. In doing so, GPT-3 transports a form of life. But the rather static word embeddings do not account for situated use. In this sense, GPT-3 de-contextualises language leading to static ideals and assumptions, which, as I will show in the following chapter, are increasingly difficult to negotiate.

In the end, the self, language and GPT-3 are entangled and co-constitute each other (see Coeckelbergh, 2018, p. 1513).

3.2 Conclusion

Ludwig Wittgenstein's notion of *language games* shows that language is interwoven with its use. This also means that language constitutes actions and thoughts of the self.

In view of contextual use, language is ambiguous. So for a language game to be meaningful, shared understanding is necessary. This shared understanding is achieved by learning many rules implicitly through the preexistence of language. As a result, language is a social and public endeavour. Moreover, language goes beyond grammatical rules that can be made explicit, as many concepts (e.g., metaphors) are based on how the self is embedded in its environment.

Overall, the self and the other are in a dynamic and reciprocal relation through language. That is, language enables interaction through which the self relates to its environment and itself. The synthetic texts by GPT-3 then shape how the self understands itself.

Chapter 4

Self-Formation in the Era of Large Language Models

In this last chapter I turn to my final research question, namely *how does the interaction with synthetic texts generated by large language models like GPT-3 debilitate indeterminate self-formation.*

As I showed in the previous chapters, the self is inherently indeterminate. GPT-3, on the other hand, generates language that conveys pre-determined ideals and identities. This is crucial because language constitutes practices and activities (i.e., interactions) and thus also the relation of the self to others and to itself. Accordingly, GPT-3 plays an important role in self-formation.

Since the self is embedded in a social context, it always faces a tension between being indeterminate and determined. In this tension, it is pivotal that the self can identify with assigned ideals or identities and negotiate them accordingly. The self or will to power organisation is then able to engage in the power struggle it faces and to channel its chaos or tension in a directed manner.

GPT-3, however, intensifies the tension between having the self formed by others (i.e., to be determined) and being an activity of the self (i.e., to be indeterminate). This is because the ideals that 'society' finds admirable are set by a comparatively small and homogeneous group of people. Through

its widespread use in various contexts, GPT-3 transports these values in a systematic, automated and accelerated manner. Moreover, as I will show, it becomes increasingly difficult for the self to negotiate the meaning of the synthetic texts.

As I will argue, GPT-3 debilitates self-formation because it (1) contains a static representation of both, language, and thus of ideals and beliefs, and the self; (2) it becomes increasingly difficult to negotiate the meaning of these representations; and (3) GPT-3 affirms the status quo by generating similar synthetic data.

It should be noted that GPT-3 is not the only will to power organisation that forms the self. Thus, despite the challenges, it does not debilitate indeterminate self-formation completely. But with language being a public endeavour and the increasing trend of LLMs, self-formation and the negotiation of societal ideals becomes more difficult. Hence, the importance of GPT-3 in the process of self-formation should not be neglected. In the concluding chapter, I will briefly make some recommendations to address the challenges.

4.1 Tension between Determinacy and Indeterminacy

Despite the indeterminacy of the self, it faces a tension between being determined and being indeterminate. This is because the self is always embedded in a social context in which ideals and values exist. For Nietzsche 'all reality is will to power' (Aydin, 2021, p. 42). And since power is inherently relational, it is impossible to step out of the game of power. Similarly, the rules of a language game already exist. Even if we were to try to create our own private language, we would need symbols available to us to do so. This means, language is inherently social and public. So it is also impossible to step out of the language game. In view of this, self-formation is not a completely autonomous process (see Aydin, 2021, pp. 147–150). As a

consequence, 'personal identity is [...] never completely "personal"' (Aydin, 2017, p. 127).

This means, others, including technology, have an image or representation of the self. In this, others should not be understood as 'external' from the self. This would presuppose a stable and independent self that revives the essentialst and dualist self. Instead, the self and others are interconnected through interaction. Others become part of the self, so to speak.

The image others have of the self can however never be the 'true' self. This is because the self lacks a pre-established essence. It is thus impossible to fully define it. Accordingly, the self's own understanding is also never complete. Rather, the representations others and the self have function as a resource for self-formation. That is, the self is defined by others, but the self also defines itself by identifying with and negotiating these definitions and beliefs. Ideals or identities are ultimately not a priori established, as suggested by Aristotle, but emerge through the interaction.

In this discrepancy between the self being determined by existing ideals and its indeterminacy, neither of these two poles should predominate too much. For a strong will to power organisation (i.e., the self), the challenge is to maintain a balance between being well organised (i.e., a seemingly stable unity) and to sense a tension or chaos in order to grow (i.e., the urge to overcome the current self). If the self were completely indeterminate, it could not form a unity because it would lack identifying boundaries and would thus fall apart. If the self, were completely determined, this would undermine its agency and 'own' control. Either because the tension or chaos (i.e., the determining force) is too great which the self is unable to organise, or because any chaos and thus struggle is cancelled out by the affirmation of existing beliefs. Both lead to decay.

Overall, then, the problem is not (primarily) to be defined by others, but rather the ability of the self to define itself in relation to the images others have of it. In other words, to negotiate these images. As Haste (2004) notes:

> Rather than being regarded as passively "socialized," the indi-

> vidual actively constructs—and co-constructs with others—explanations and stories that make sense of experience, to develop an identity that locates her or him in a social, cultural, and historical context. (p. 420)

Self-formation is thus a deliberate, reflective and conscious process, that allows for negotiating the meaning of certain ideals or assigned identities. This, in turn, allows to overcome or reinterpret past behaviour and in doing so to (radically) change. The goal, then, is to value ambiguity and pluralism that allow for alternative ways of being, instead of adhering to 'standardised' ideals.

As I will show in the following section, it is increasingly difficult, if not impossible, for the self to co-construct and (effectively) negotiate these ideals, or explanations and stories embedded in GPT-3. This does not mean that the synthetic texts cannot be contested or negotiated per se, but rather the cause, i.e., GPT-3. As a result, GPT-3 intensifies the challenge of having the self formed by others (i.e., to be determined) and forming the self on its own (i.e., to be indeterminate). It thus debilitates self-formation.

4.2 The Challenges that GPT-3 poses for Self-Formation

So far I have treated the self as a generic entity. When it comes to the interaction with GPT-3, however, there are many different selves to consider. For example, someone who constructs the material infrastructure, builds the model, uses it to produce other applications, or consumes the content generated. Given this complexity, I will mainly focus on self-formation of a self that interacts with the output of GPT-3 (i.e., the consumption of synthetic texts). Hence, the self does not interact directly with GPT-3, but through other services, such as a chat bot, news site, or a search engine, that make use of the language model. In a sense, GPT-3 takes the place of the otherwise

human interlocutor.

At this point, I want to establish the notion of GPT-3 as a will-to-power organisation.[1] In doing so, I do not want to attribute any conscious will to LLMs. It might be more appropriate to understand GPT-3 as the bearer, intermediate or extension of other will to power organisations (e.g., OpenAI, Microsoft or the ideals reflected in the data sources like Reddit) that result from socio-technical practises (see Seaver, 2017). For simplicity, however, I will refer to GPT-3 as a will to power organisation. Besides, as Lash (2007) claims:

> A society of ubiquitous media means a society in which power is increasingly in the algorithm (p. 71).

Framing GPT-3 as a will to power organisation allows me to conceptualise the interaction between the self and GPT-3 as a power struggle and thus as negotiation.

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

---

[1] Although GPT-3 possibly debilitates self-formation and thus seems 'stronger', it can nevertheless be understood as a weak or sick will to power organisation as it does not possess any struggle or chaos. On the contrary, its mathematical optimisation functions strive for standardisation and normalisation.

4.2.1 Static and Reductionist Representation

Of Language

Through the notion of Wittgenstein's language games, we see that language is a product whose meaning emerges through use. To understand the meaning of words or jargon (i.e., signifier) that contain domain-specific concepts (i.e., signified), we need to consider context (M. Mitchell, 2020, p. 226). That is, who says something, in which situation, with what kind of intonation and thus with what intention. So, not only quantitative aspects matter, but also qualitative ones. Thereby, qualitative features are based on shared understanding and tacit knowledge, which allows to resolve (or reduce) the ambiguity of language. As such, the word 'nigger', for example, can have 'various positive in-group meanings and pejorative out-group meanings' (Alim, 2006, p. 77). Consequently, meaning is not reducible to words, which makes language inherently contextual and reciprocal. Hence, language cannot be decoupled from use.

Through statistical correlations and probabilistic calculations of word vectors, GPT-3 reduces words to numerical values.[2] In doing so, qualitative and contextual factors are removed (e.g., Bisk et al., 2020).[3] The result is a static representation of language stemming from a very specific context. Although the goal of LLMs might not be to adequately represent the entirety of lan-

[2]Crawford (2021, pp. 89–95) makes a similar point with the example of mug shots. The detainees and the pictures of them are merely considered as data points in various databases, without taking into account the contextual background history of these people.

[3]GPT-3 might contain more contextual information compared to previous models due to the increase in computational resources and the amount of data processed. Nevertheless, I 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 (M. Mitchell, 2020, p. 267). See also Winograd (1990, p. 179).

guage, GPT-3 is nevertheless applied across various contexts. Besides, as Forsythe (1993) describes the early expert systems:

> 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)

So then the problem with GPT-3 lies in modifying and negotiating that (static) representation. I will come back to that in a moment (section 4.2.2). Especially considering the contextual and situated use of language. In the end, the formalisation of language, creates the illusion that language or words can stand for themselves, independent of its use.

This illusion of independence might remind us of Descartes' dualism. That is, we should not be deceived by our senses (i.e., the outside) to arrive at objective knowledge (see Birhane, 2021, p. 3). The assumption is that information simply travels from the outside to the inside where it can be scrutinised accordingly. In the case of GPT-3, data (the outside) is gathered and manipulated by mathematical operations (the inside). And in view of the large corpus of text GPT-3 was trained on, it might appear that GPT-3 learns in a more neutral way. But as many examples illustrate, GPT-3 produces value-laden texts. Consequently, the de-coupling of inside and outside, of words and their use, arouses a false image of objectivity (see Husserl, 1970, p. 79).

This is significant insofar as:

> Language ideologies play a vital role in reinforcing and justifying social hierarchies because beliefs about language varieties or practices often translate into beliefs about their speakers. (Blodgett et al., 2020, p. 5459)

For GPT-3 it is the other way around. It adopts and perpetuates the language ideologies of the 'speakers' (i.e., mainly young white males) reflected in the

training data scraped from Internet sources. The de-contextualisation of language thus leads to GPT-3 conveying a comparatively narrow worldview and ultimately a particular view of the self (see Aydin & de Boer, 2020, p. 732). In terms of Ludwig Wittgenstein, GPT-3 conveys a form of life including certain beliefs and assumptions through the language it generates.

Of the Self

In general, machine learning models applied in a social context are based on the assumption that the self can be defined and categorised and in turn predicted. Be it by race, gender, hobbies, political or religious beliefs. As such, questionable applications claim to be able to infer sexual (see Sullivan, 2019, p. 21) or political orientation (Kosinski, 2021) from facial features. Jones (2018) states that:

> Much of the promise of the data sciences, whether in medicine, marketing, or getting out the vote, ostensibly comes from overcoming older theory-laden categorizations to characterize individuals in their specificity, all in order to predict their behavior. (p. 684)

Similar to language, the self is not static, but inherently dynamic and in the process of becoming. In this, the self has no pre-established identity and can thus never be fully defined. It can be characterised as parent, student, tennis player and so forth. And yet these identities, even in sum, do not describe the self in its entirety. Because these identities are only a temporarily result of organising different wills to power through interaction, which results in a seemingly stable will to power organisation (like a system). Put differently, these identities can overlap and change from one instance to the other. Consequently, its indeterminate nature make the self unpredictable. In a sense, there is no concrete or complete signified (i.e., concept) of the self.

GPT-3 does not (by itself) categorise the self or predict a particular identity based on past behaviour. But it nonetheless generates narratives that

contain certain beliefs and identities. Such as associating certain professions with gender, or certain races with lower sentiment. In the end, GPT-3 does not account for the fluidity and plurality of the self. Instead, by reducing the ambiguity of language, GPT-3 generalises and reduces the self. This 'normalisation' or 'standardisation' of the self undermines its indeterminacy in that it excludes other alternatives.

Certainly, the self does not have to identify with these assigned categories. As Aydin (2021) notes:

> The homogenizing of reality in this way does not lead to the negation of the vitality, diversity and richness of the world. On the contrary, due to it, every determination of reality, every interpretation, can be continuously questioned by opposing powers; because of this, other interpretations always remain possible. (p. 44)

A survey of 963 Facebook users, for example, shows that 260 participants (27%) disagree with the labels assigned by the platform, while 491 participants (51%) are uncomfortable being categorised (Hitlin & Rainie, 2019, p. 7). The dissatisfaction thus 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. Further, those affected, often already marginalised in society, have to live with the (often harmful) consequences, such as predictive policing leading to false arrests or GPT-3 creating bigoted texts. Besides, although 568 out of the 963 participants (59%) think Facebook categorises them correctly, they do not know why that is the case. This brings me to my next point.

4.2.2 Unidirectional Interaction

As mentioned, there are various different selves or will to power organisations that interact with GPT-3. Some might be aware of the implications of GPT-3, but lack the power or know-how to act accordingly (as an individual). Still

others may not know about the existence of GPT-3. The question then arises as to who has the power to challenge and negotiate the values embedded in GPT-3. Although the following will not lead to a breakdown of different stakeholders, I want to argue that the invisibility and incomprehensibility of GPT-3 undermines the possibility to modify and negotiate the embedded worldview.

Invisibility of GPT-3

In general, the self 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 the meaning-making. Accordingly, self has the possibility to challenge these beliefs or assigned identities (more easily). This is because, first, an interaction or dialogue between people is a constant back-and-forth and (theoretically) open-ended interaction. This reciprocity makes it easier to question, contest or refute certain views. In this way, assumptions can be re-interpreted and re-evaluated. Second, since a dialogue takes place in a shared and situated context, it is clear who is participating in the language game and what rules are to be followed accordingly. This in turn allows the will to power organisation (i.e., the self) organise itself based on context. These characteristics make a dialogue less rigid or static leaving room for (contextual) ambiguity and negotiation. In other words, a dialogue is (more or less) indeterminate.

This informed and reflective interaction or dialogue with GPT-3 is very difficult, if not impossible. For various machine learning models the self may not be aware that it is interacting with them, and more specifically what that interaction entails. For example, many American students were unaware of the fact that their Facebook feed is algorithmically sorted (Powers, 2017). Another survey shows that 712 out 963 (74%) Facebook users were unaware that the platform categorises them based on assumed interests (Hitlin & Rainie, 2019). The texts of GPT-3 are also difficult to distinguish, which is

why it is not clear whether the self is reading an article written by a machine that is tailored to its interests or not (unless it is disclosed). These issues might raise concerns about trust and responsibility. But the point is that the seamless nature of LLMs makes them 'invisible active actor[s]' (Beerends and Aydin, 2021) in the process of self-formation.

The invisibility[4] or unawareness creates an illusion of independence, which in turn creates a false impression of autonomy of the self. That is, GPT-3 has supposedly no effect on the self. But since 'power is only power in relation to another power' (Aydin, 2021, p. 42), power is inherently relational. This means we cannot step out of the game of power. Consequently, if the self is not aware of who or what it is interacting with at any given moment, power is not absent, but concealed.

This concealment leads to the self being unable to (effectively) organise the power struggle, because it cannot be directed against the originator. As a consequence, the interaction with these models is not reciprocal, but unidirectional. Even if the self wants to refrain from the interaction with GPT-3, it has to be aware of its existence in the first place. Besides, the ubiquity of the model makes it increasingly difficult to avoid interaction. Overall, this makes deliberate and reflective formation of the self difficult.

Authors who deliberately interact with GPT-3 can, of course, modify the text produced by GPT-3. Likewise, suggested translations can be adjusted. But as studies show, machines can influence the choice of words (Brandstetter & Bartneck, 2017, p. 284). Besides, it becomes more difficult to modify the output in cases where GPT-3 processes search engine queries (Metzler et al., 2021) or converses with the self in the form of a chat bot. But even if the self does not coincide with the generated values or identities, or changes a given text, this does not stop GPT-3 from continuing to produce these narratives. Rather, GPT-3 will maintain the same assumptions.

[4]Machine learning models are not invisible in the sense of being immaterial. Their infrastructure is heavily dependent on physical resources (see Ensmenger, 2013, 2018).

This is important because it is very difficult to question why a particular output was generated. In other words, the self does not know how meaning is derived. As Beerends and Aydin (2021, p. 1676) put it:

> The authenticity of algorithmically assumed characteristics and affinity categories cannot be negotiated as users have no access to the meanings derived from their data and how this is fed back to them.

Incomprehensibility of GPT-3

Even for developers and engineers, it becomes increasingly difficult to conclude why an algorithm has drawn a certain conclusion. Usually, rule-based systems, such as the ELIZA chat bot, are comprehensible as they rely on human knowledge. Subsymbolic or deep learning systems, however, are not based on human representation of the world, but learn correlations themselves. The complexity of these models make them increasingly opaque. This in turn leads to causal uncertainty in the way assumptions are used to generate an output. An image recognition system, for example, (supposedly) discerned friendly tanks from enemies based on daylight (see Zednik, 2019). While this inadequate correlation was detected, we may not (yet) comprehend the more subtle assumptions embedded in large language models (e.g., Blodgett et al., 2020, p. 5460). Also because language is contextual and stereotypes or other prejudices thus vary (Weidinger et al., 2021, p. 12). Independently, for example, a medical chat bot run by GPT-3, suggested starting to recycle to overcome sadness or even commit suicide (Rousseau et al., 2020).

In the end, these subsymbolic systems (i.e., LLMs) do not represent the way we humans think and how we perceive the world. So although the model has a representation of language, it is not based on ours. This is because GPT-3 simply correlates the occurrences of words. In that sense, LLMs are incomprehensible because of a lack of shared understanding. Even despite the great efforts in explainable and interpretable AI (e.g., Doshi-Velez & Kim,

2017; Schramowski et al., 2021), the problem of complexity of the underlying functioning remains. Besides, explanations can sometimes be misleading (Fischer, 2020; Rudin, 2019). Given this, engineers and developers are not (sufficiently) capable of engaging in a 'reflective dialogue' with GPT-3 to question its embedded ideals and beliefs. Hence, the interaction is again not reciprocal, but unidirectional.

Again, the self can still reject the output by GPT-3. Even in the case where the medical chat bot suggested committing suicide, the self does not have to follow this suggestion. Whether it is the responsibility of a (mentally unstable) patient to judge whether the suggestion is appropriate or not is further questionable. And even if it were the case that a therapist suggested suicide, it is very likely that a different outcome would follow. This is because instead of having to cope with the suggestion themself, as is the case with the chat bot, the patient can turn their attention to the therapist and respond. In a sense, the self can organise its struggle and direct its tension specifically to the cause. Consequently, it depends on the (situated and shared) context that shapes the interaction (e.g., De Jaegher & Di Paolo, 2007), and in which the self has the possibility to negotiate suggestions or assigned characteristics accordingly.

Overall, the invisibility or unawareness and incomprehensibility of GPT-3 undermine the possibility to question and challenge embedded assumptions. This is because the self is unable to organise its power struggle. In other words, the self cannot (effectively) negotiate the meaning generated. As a result, the self becomes a patient in a unidirectional interaction (see Franklin, 1990, pp. 48–49) in which GPT-3 forms the self, but it is difficult, almost impossible, for the self to shape GPT-3. This is crucial because a will to power organisation that is or cannot be contested might be accepted as ground truth and ultimately as reality (Aydin, 2007, p. 36).[5]

[5]This should not be understood as some inevitable technological determinism. The self still can change the functioning of GPT-3. I will elaborate on this in the final conclusion.

4.2.3 Reinforcing the Status Quo

The undefined and thus unfinished nature of the self leaves it in a process of becoming, whereby it can (radically) change:

> Change can be the symptom of both the establishment of a new hierarchical order ("substantializing") as well as the collapse of an old order ("de-substantializing"). (Aydin, 2021, p. 44)

It seems, however, as if GPT-3 neither establishes a new hierarchical order, nor collapses an old one. As many other machine learning models, GPT-3 rests on the assumption that the future resembles the past. That is, past events or behaviour are quantified. This data is then used to make predictions about the future.

In certain use cases where the problem space is limited and well explored, such as chess or skin cancer detection in medical images[6] (see Sullivan, 2019, p. 20), this static assumption might be appropriate and even necessary. In these cases, the relationship between cause and effect is unlikely to change and 'the model is operating within a background of existing scientific understanding' (Sullivan, 2019, p. 20). For both language and the self, however, this (scientific) certainty or stability is not given. Because both are inherently dynamic, contextual, unfinished and thus unpredictable.

Again, GPT-3 does not attempt to predict the essence of the self, nor the totality of language, but it nonetheless conveys a form of life containing beliefs, ideals and identities. As shown, the results include texts that make undesirable connections with mentions of disability (Hutchinson et al., 2020), that present various gender stereotypes (Lucy & Bamman, 2021), or that assume that a family consists (exclusively) of a married women and man with children (Weidinger et al., 2021, p. 13). GPT-3 is ultimately a 'stochastic

[6]This is not to say that these systems work without any problems as they can still contain representation biases that result in poor performance with darker skin tones (e.g., Guo et al., 2021).

parrot' that recycles old beliefs and a particular worldview (Bender et al., 2021, pp. 613–5).

The reason is that GPT-3 generative capabilities depend on the data distribution which it derives from historical data through probabilistic calculations. Hence, the emergence of the generated output is already existent in the data set, instead of being (radically) new. This does not mean that generative models repeat the same output over and over again. Rather, the output is almost identical to the input (e.g., Abid et al., 2021). This behaviour is not unique to GPT-3. The project 'This person does not exist'[7], for example, creates images of faces through a so-called generative adversarial network (see Goodfellow et al., 2014). Researchers, however, found that the generated images resemble to a high degree the faces in the training data (Webster et al., 2021). Another example is Facebook's model that links nursing jobs to women and technical jobs to men and displays job ads accordingly (Lambrecht & Tucker, 2019).

The same might be said about the self as its ideals or its language do not emerge out of nowhere. The self, however, is more fluid and more likely to change. Machine learning models, in contrast, are more rigid and do not account for temporal changes accordingly. Thus, '[t]ransformer [L]LMs become increasingly outdated with time' (Lazaridou et al., 2021, p. 9). This is because, to learn new correlations, the model requires thousands of examples (which may not yet be available), whereas a comparatively small amount is often sufficient for the self. This is crucial because the self can never be completely defined. Rigidity thus constraints self-formation by hindering the overcoming of the present state.

Perhaps it can be argued that machine learning models allows the self to challenge the status quo in that they reveal current power structures and beliefs held in certain communities (see Srinivasan, 2012, p. 206). Perhaps otherwise these patterns would have remained invisible. Large language mod-

[7]https://thispersondoesnotexist.com/

els, in a sense, make explicit the tacit rules of the language game being played in a particular form of life. This, in turn, allows the self to contrast them to other ideals and overcome them accordingly. This apparent 'advantage', however, should not be understood as justification of the social biases embedded in GPT-3. This is because those already marginalised by society have to live with the harmful consequences. So, while GPT-3 may mean desirable self-formation for some, it is detrimental to many others. As Hong (2021) notes:

> What is at stake in this repetition and narrowing is not simply what kinds of futures can be "imagined," but which futures – and *whose* futures – are prioritized with unerring regularity at the expense of which others. (p. 1942)

Feedback Loops

Repetition and the danger of 'value-lock' (Bender et al., 2021, p. 614) is further strengthened by feedback loops.

Although I framed the interaction between the self and GPT-3 as unidirectional, this is not entirely true. Because during the interaction data is generated. In addition to presumably behavioural data, primarily synthetic output from GPT-3. This data is fed back to the system at some point, which in turn shapes the next interaction between the self and GPT-3. On that note, one may think of so-called *filter bubbles* which search engines or social media sites create based on past data. YouTube, for example, recommends videos based on our viewing history. Researchers also fear the spread of misinformation and thus radicalisation through the use of GPT-3 (McGuffie & Newhouse, 2020). This is because already existing assumptions or beliefs could be further reinforced.

As presented by O'Neil (2016), several machine learning models create feedback loops. For example, in the case of predictive policing: Based on historic data that was used to train the model, it predicts a crime, which the officer follows. If an arrest is made, it leads to more data which is fed

back to the system and confirms the prediction. If no arrest is made, no data is generated, hence there is no information to correct the false assumption of the model. As a consequence, there is a higher tendency to associate a criminal identity with categories such as 'Black' or 'Hispanic' (Mohler et al., 2018, p. 2457).

The feedback, and especially the lack of it, thus ultimately optimises the predetermined target function and 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'. In other words, the (lack of) feedback 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. And as a result, these models create an environment that justifies the initial assumptions.

Moreover, GPT-3 might feed its own confirmation bias, as synthetic data very likely finds its way into future training data used for subsequent LLMs. The problem that both training data and test data scraped from the Internet can overlap is also noted by the OpenAI researchers (Brown et al., 2020, p. 29). Accordingly, Raji et al. (2021) also argue that benchmark tests on the performance of a model can be misleading. Under these considerations, GPT-3 perpetuates past behaviour and ultimately creates a feedback loop – for itself and for the self (e.g., Weidinger et al., 2021, p. 14).

In doing so, it maintains and perpetuates existing assumptions and power structures (Blodgett et al., 2020). Put differently, GPT-3 preserves a certain kind of life (see Aydin, 2007, p. 37). Conversely, GPT-3 neither establishes a new hierarchical order, nor does it collapse an old one.

Reviving the Past

Machine learning models perpetuate past behaviour in a subtle, yet systematic manner. In a Wired article, Lauren Goode (2021) tells how the Memories function of her photo app kept reminding her of her wedding that she had cancelled. Not only did the photo app keep highlighting photos of her and

her fiance, but different websites displayed advertisements of wedding dresses, and several services sent her email reminders. For the latter it is most often possible to unsubscribe. But for the apps that are run by machine learning models, it is much more difficult if not even impossible.

The process of training a new model is very time-consuming. First, the problem must be identified, then a solution must be found for updating the model or data, or both. If neither can be adjusted appropriately, new data must be collected that corrects the behaviour. Provided that this data exists in the first place (Weidinger et al., 2021, p. 12). Overall, this process involves high compute costs, both financial and environmental (Bender et al., 2021, p. 614).[8] Accordingly, this raises the problem of how to modify and unlearn these systems.

The self is without doubt a historical being. The difference, however, is to be determined by a past or by 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 in 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). Thereby, forgetting is not meant as merely ignoring or erasing the past, but to overcome certain ideals or events by re-interpreting and re-negotiating them (see Kudina, 2022).

In her book, Kate Eichhorn (2019) discusses the consequences of young adolescents growing up in in the era of social media, which introduces 'the end of forgetting'. As she mentions, (digital) memories can no longer be easily modified, and given the ubiquity

> the past can now more easily seep into the present. Neither space nor time serves as a significant barrier. (p. 142)

As a result, the self is limited in its ability to explore different identities. Similarly, GPT-3, revives the past by, for example, neglecting gender-neutral

[8]The computational resources required for a training run add up to USD 4.6 million (Dale, 2021, p. 115)

pronouns, thereby excluding alternative identities (Weidinger et al., 2021, pp. 13–4). Similar to rigidity, GPT-3, by reviving past ideals or assumptions, again prevents the self from overcoming its current state.

Chaos and Alternative Worlds

Despite the many examples that contain negative consequences, feedback loops also apply to 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). In the case of GPT-3, it would be possible to produce personalised stories with other specific (behavioural) data points. So the self might think it acts by its own will, but instead, ideals it might have adopted from others are not questioned, but reinforced. The problem with this is that, as Nietzsche (1997) notes when discussing 'own evaluations' and 'opinions':

> 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, p. 61)

The self is, as mentioned earlier, always a social and a self-made product. Hence, the self cannot have a completely 'personal' opinion. Nevertheless, through personalised stories values are transported in a very subtle way, which affirm currently held ideals. This makes it increasingly difficult to contrast these ideals, as it leaves out other alternatives (see Aydin, 2021, p. 48). Moreover, the way language transports ideals can itself be very subtle. For example, word order can maintain power hierarchies (e.g., woman and man vs. man and woman) (Mooney & Evans, 2015, p. 112). Ultimately, by creating these feedback loops that affirm the status quo, GPT-3 reduces the unknown or the 'chaos':

> There is what he [Nietzsche] sometimes calls a 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. (Aydin, 2021, pp. 43–44)

But by quantifying, sorting and categorising language, and inferring old beliefs GPT-3 suppresses this chaos. The reduction of chaos, however, renders the will to power organisation (i.e., the self) weak. As struggle is reduced, but struggle is necessary for growth.

The de-contextualisation and thus reductionist representation of language eliminates ambiguity which leads to a static view of the self. And since the meaning of synthetic texts cannot be negotiated, the status quo is affirmed, which in turn reduces plurality and other alternatives. As a result, ideals become more and more conformist, homogeneous and ultimately standardised, which again excludes alternative worlds, i.e., new ways of being, and so on. Ultimately, this undermines the indeterminacy of the self (see Aydin, 2021, pp. 173–4).

4.2.4 Uniformity

So far I might have idealised and attributed too much capabilities to GPT-3. But GPT-3 has its limitations. For example, it loses coherence the longer a text becomes or it has troubles giving answers to – for us – simple questions (Elkins & Chun, 2020, p. 3).[9] Further, GPT-3 is not the only will to power organisation that forms the self, rather there are various other (non-technical) will to power organisations. Thus, I do not want to claim that GPT-3 completely debilitates indeterminate self-formation per se.

[9]Similarly, the chat bot ELIZA was also thought to behave more intelligently than it actually did (Bender & Koller, 2020, p. 5188). In this context, one might also think of the horse 'Clever Hans', which supposedly could perform mathematical calculations (Crawford, 2021, pp. 1–7).

Nevertheless, this does not mean that we should dismiss the importance of GPT-3 it has on self-formation. Especially since GPT-3 differs from other models in that it generates language, the basis of human communication.Besides, there is an increasing trend in recent years to ever larger language models that sift through larger amounts of data (Brown et al., 2020, p. 4). As researchers from Stanford University claim, we are entering a new paradigm through what they call *Foundation Models* (Bommasani et al., 2021). As they open their report:

> AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. [...] Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization. (p.1)

So GPT-3 may not have a direct impact on any particular self, but it will very likely have an impact on society at large through growing acceptance and use. As such, societal ideals are shaped by GPT-3. And since the self is always embedded in a social context, it has to identify in one way or another with these ideals that exist in society. In this case, GPT-3 affects self-formation in an indirect way.

Further, GPT-3 is not a single technology, but enables various services and applications. Nevertheless, applications run by GPT-3 are based on the same linguistic representation. So even if it is possible not to interact with one of these services, or to prohibit them, the existence of GPT-3 and LLMs in general is likely to remain. Especially given Microsoft's involvement, makes it likely that the GPT-3 will become part of our lives in one way or another, just like Microsoft itself. Similarly, socio-economic dynamics make it likely that news media organisations, for example, could demand the use of GPT-3. This would affect the production of texts, which in turn affects

consumption, and vice versa (e.g., S. Bishop, 2018; Caplan & boyd danah, 2018).

As we can see, GPT-3 not only contributes to a homogenisation of ideals and values that influence thoughts or beliefs of the self (i.e., on a micro-level), but also to a larger paradigm (i.e., on a meso/macro-level) that systematically suppresses chaos. This conformity or uniformity was also expressed by Nietzsche:

> he believes that in his (our?) culture, Platonic metaphysics and Christian morality have organized life in a uniform way to such a degree that every possible counteraction is destroyed. (Aydin, 2021, p. 48)

Applied to our (technological) culture, one could say that LLMs (and other machine learning systems) try to organise our lives in a uniform way (see Sutherland, 2014, p. 547). In the process, 'counteraction is destroyed' by the inability to negotiate the meaning and by the overarching societal acceptance and aspiration of quantification. Goals like 'I think you can capture a thought by a vector' expressed by Geoffrey Hinton from Google or 'we want to embed the world in thought vectors. We call this *World2Vec*' stated by Yann LeCun from Facebook (M. Mitchell, 2020, p. 250), further exemplify how categorising, ordering and thus striving for stability in every aspect of life is the ultimate perspective.

But Nietzsche (1974) warns against putting too much faith in the natural sciences, which seeks a 'world of truth' based on mere reasoning and measurements:

> Do we really want to permit existence to be degraded for us like this – reduced to a mere exercise for a calculator and an indoor diversion for mathematicians? Above all, one should not wish to divest existence of its *rich ambiguity*. (§373, p. 335)

Since the self is indeterminate and can thus never be adequately defined,

I want to advocate for appreciating ambiguity, diversity, plurality and ultimately chaos. For:

> [t]he more that chaos breaks into our ordered world, the more our creative power is stimulated. (Aydin, 2021, p. 44)

This creative power allows for different ways of being and the exploration of alternative worlds that go beyond categories that can be captured by numerical values. Ultimately, ambiguity reduces the constraints of indeterminate self-formation.

4.3 Conclusion

In this last chapter I addressed my research question: *How does the interaction with synthetic texts generated by large language models like GPT-3 debilitate indeterminate self-formation.*

In general, the self always finds itself in a tension between being determined and being indeterminate. This is because, the self is embedded in a societal context, in which certain ideals and beliefs already preexist. The ability to negotiate these ideals or assigned identities are thus the precondition for self-formation. The self is ultimately a social and self-made product.

Institutions or other people also assign identities to the self and uphold certain ideals. Today, however, many deep learning models perform this task. What distinguished these models is their ubiquity, speed, and (apparent) specificity. Crucially, what 'society' finds admirable is thereby defined by a comparatively small and homogeneous group of people who develop these systems. Accordingly, certain values are transported in a systematic, accelerated and automated manner across contexts.

With Wittgenstein, we see that language constitutes the practices and activities of the self. In this, both language and the self are inherently dynamic, contextual, and ambiguous. GPT-3, on the other hand, holds a static representation of language. This, in turn, leads to a static and pre-established

identity of the self. This is because language embodies beliefs and a form of life. As a consequence, the self is reduced to a limited (and prejudiced) identity, negating its plurality. The reduction to a particular identity is contrary to Nietzsche's understanding that the self has no essence. Further, the assumption that language is independent of use can be related to Descartes' dualism, where information is considered as something that merely travels from the inside to the outside. But since language is highly contextual, decoupling language and from its use leads to a false sense of objectivity.

To engage in (an equal) power struggle or (effective) negotiation, a reciprocal interaction is necessary. The interaction with GPT-3, however, is unidirectional. This is because the self is often unaware that it even interacts with the language model, which undermines a deliberate negotiation. Accordingly, a false sense of autonomy might arise. But since power is always relational, the power of GPT-3 is not absent, but obscured. Further, due to the complexity and incomprehensibility of GPT-3, even engineers have difficulties scrutinising or changing built-in assumptions. So again, the interaction is not (completely) reciprocal. Hence, the unidirectional character undermines the possibility of a power struggle and thus the possibility to negotiate the meaning of beliefs or ascribed identities.

Most importantly, the self is inherently indeterminate and thus unpredictable. This, however, goes against the nature of machine learning models that are based on calculating predictions based on historic data. In that way, the assumption is that change follows a pregiven path. Put differently, the future resembles the past. By recycling old beliefs, GPT-3 affirms the status quo. And since the unidirectional character undermines the possibility to negotiate the pregiven, static representations, the interaction between GPT-3 and the self creates a feedback loop. This feedback loop, in turn, reinforces the status quo. In doing so, the self cannot overcome its current state.

Overall, the mathematical formalisation of language, with its quantification, sorting and categorising, reduces the measureless variety and multiplicity of possibilities, i.e., the chaos. This chaos, however, is necessary to

discover alternative worlds. This applies to the individual self as well as to the political or societal level. So, while GPT-3 is of course not the only will to power that forms the self, the use of large language models is likely to increase nonetheless. And with it, the importance that these models have in the process of self-formation. The self is always embedded in an environment, and as Wittgenstein notes, there is no such thing as a private language. Self-formation, then, is also a societal endeavour.

Given the above, the answer to my research question is that GPT-3 debilitates indeterminate self-formation because it undermines a deliberate and reflective process in which negotiating certain ideals or imposed identities is possible. And in combination with the re-use of old assumptions, GPT-3 undermines the possibility of overcoming past ideals, which prevents alternative ways of being and thus (radical) change.

Conclusion

Before I move to the final conclusion, I want to point out some limitations and provide a brief outlook for further research.

Limitations

I emphasised the importance of context and yet I did not consider a specific context. Thus, it might be said that my considerations are perhaps too broad or general. It is because, first, GPT-3 is not a single model or technology. Rather, GPT-3 enables various different services and applications. For example, a medical chat bot poses different challenges to self-formation, than a bot that posts cricket results on Twitter. Accordingly, GPT-3 affects different practices, whether social, political, or economic in varying degrees. Second, the self is also a generic concept. Each interaction between the self and GPT-3 is different and thus unique. Hence, each self will experience the outcome of the interaction differently. In other words, self-formation means different things to different selves. To overcome these limitations, further research is necessary that takes into account the lived experiences of selves with a particular application applied to a particular context (see Blodgett et al., 2020, p. 5458).

I nevertheless think the underlying problem remains. Namely, that the self is not able to negotiate the meaning of the output generated, including the assumptions that lead to it. By re-recycling old assumptions and ideals other alternatives are excluded. This, in turn, reinforces current power struc-

tures, i.e., the status quo, contributing to homogeneous ideals. Consequently, GPT-3 increases the tension between 'hetero-formation' by others and indeterminate self-formation (i.e., activity of the self), especially for those who are already marginalised from society.

Having said that, the main focus should not be on GPT only. So although I framed GPT-3 as a will to power organisation, it should, as also mentioned, rather be understood as the bearer of values or an extension of other will to power organisations (e.g., data sources, Mircosoft, OpenAI). Thus, further research and solutions are needed to enable a more open and pluralistic society that promotes different ways of being.

Outlook

Despite the many challenges that GPT-3 poses, we can still change the model and the interactions we have with it. This is because, the construction and validation of a large language model is only one possibility of many others (see Raji et al., 2021, p. 6).

In order to allow for ambiguity, we need to take context into account (M. Mitchell et al., 2019; Raji et al., 2021, p. 7).[10] This means we need to reconsider how we collect and annotate data (e.g., Gebru et al., 2018; Jo & Gebru, 2020). Boumans and Leonelli (2020, p. 96), for example, stress the importance of defining the 'richness, variability and multiplicity of features' in plant phenomics. Otherwise, without context-dependent information, the data cannot be situated and thus loses its usefulness outside of the context it was collected. Leonelli (2014, p. 8) also notes that we need to

> disaggregate the notion of Big Data science as a homogeneous whole and instead pay attention to its specific manifestations across different contexts.

[10]Inspired by M. Mitchell et al. (2019), GPT-3 provides a model card. See https://github.com/openai/gpt-3/blob/master/model-card.md

Despite context specificity, however, there is no data set that is neutral or universal (Raji et al., 2021, p. 8). In the end, '[a]ll researchers are interpreters of data' (boyd danah and Crawford, 2012, p. 667). Consequently, we need to acknowledge that big data or data science is not objective or neutral.

Often it is suggested to 'unbias' data, but in some cases 'bias' can be useful. For example, a model that contains prejudices or slurs could be used for content moderation and filter out such language. This is, of course, a highly normative task. But the point is that the same data can lead to different results. So even bias is contextual.

Nevertheless, 'addressing bias in a dataset is a tiny Band-Aid for a much larger problem' (D'Ignazio and Klein, 2020, p. 60). Because in doing so, the more deep rooted social inequalities are overlooked (Birhane, 2021, p. 2). So in the end, it is not a question about bias, but about power (Miceli et al., 2022).

As Plasek (2016) summarises:

> What we need is a better appreciation for the deeply contingent and mutually constitutive role between computation and data and the material, political, and social circumstances that inform how these algorithms are instantiated through different communities of practice. To begin to see how biases are propagated and reinforced via machine learning system training data, we need histories of the datasets themselves: how they were formed and why, how they have been maintained and subsequently altered, and how they were valued as useful data for the problems to which they were employed. (p. 6)

Accordingly, we need processes that make visible how data was appropriated and negotiated. Given the political and economic factors of data collection and model construction (Leonelli, 2014, p. 7; Miceli et al., 2022), self-formation is a also a societal endeavour (see Aydin, 2021, p. 149). We therefore need to create an environment, in which a reflective and deliberate interaction with LLMs is possible. This requires an awareness of the

challenges and consequences that large language models bring. Further, we need to create interactions that are not unidirectional, but reciprocal. In other words, we need to create possibilities to negotiate the meaning of the generated output. This includes making visible why a model infers certain assumptions.

Ultimately, however, we need to critically reflect on the appropriateness of assuming an identity based on past behaviour in a given situation. For example, the aim for predictive policing systems is not to stop all selves equally. Further, we need to take into account the consequences of potential errors. As, for example, a medical bot suggesting suicide. Hence, we must question the usefulness of the model itself. Because, again, regardless of how much context we add to the data, or how well we design the interaction,

> there is no dataset that will be able to capture the full complexity of the details of existence. (Raji et al., 2021, p. 10)

This is because both the self and language emerge through interaction, i.e., in the unquantifiable *in-between*.

Summary

For Nietzsche, the self has no essence (i.e., identity), but is inherently indeterminate. This means, the self is a priori undefined and its development remains unfinished and unknown. In the process of becoming, the self is formed through interactions.

During interaction, the self is always confronted with ideals or assigned identities, be it by others or by technology. In view of its indeterminacy, however, the self can never be fully defined. Conversely, the self can never understand itself completely (on its own). Instead, the self forms itself through the negotiation of the ideals or identities that stand *between* the self and the other. And since the self never 'is', this is an ongoing process. In other words, self-formation is understood as constant re-negotiation of ideals or character-

istics. This, in turn, allows the self to grow by overcoming its current state and to (radically) change.

Large language models, like GPT-3, affect self-formation in that they generate synthetic texts. The self uses language to exchange beliefs or goals, which in turn shape actions. Further, through language the self relates to others and understands itself in this way. This understanding, however, is not fixed. Because language itself is contextual, situated and thus ambiguous. Accordingly, both the self and language are dynamic, fluid and open-ended.

GPT-3, in contrast, contains language stemming from a very specific context. As a result, the generated texts contain social stereotypes and prejudices related to gender, race, religion and ableism. Through its prevalence, GPT-3 conveys these particular ideals and views of self across contexts. By de-contextualising language, GPT-3 reduces and 'normalises' the self.

Ultimately, GPT-3 debilitates self-formation in that it undermines the possibility of (effectively) negotiating the meaning of the output generated. Due to the invisibility and incomprehensibility, the interaction between the self and GPT-3 is not reciprocal, but unidirectional. As a result, the underlying assumptions cannot be questioned, challenged or refuted. In case the self coincides with the ideals of the synthetic text, it affirms already held beliefs, but this limits growth. In case the self does not identify with these ideals, it cannot 'correct' or modify the meaning or underlying assumptions that lead to the output. In both cases, old ideals and thus power structures are recycled. As a result, GPT-3 creates a feedback loop and thus excludes other alternatives. This ultimately stands in the way of becoming and (radical) change.

Indeterminate self-formation nevertheless remains possible because GPT-3 is not the only will to power organisation that forms the self. The increasing trend towards more and larger language models, however, will likely affect societal ideals, which in turn affect the self. Consequently, self-formation is also a societal endeavour.

To reduce the constraints of self-formation, new ways of interacting with

GPT-3 are required. At its core, we must stop treating LLMs (and machine learning models in general) as objective and universal. Instead, we must acknowledge and appreciate the contextual and thus pluralistic and ambiguous nature of both language and the self. By securing the indeterminacy of the self, alternative and different ways of being become possible that go beyond quantifiable categories.

Bibliography

Articles and Books

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*, 298–306. https://doi.org/10.1145/3461702.3462624

Alim, S. H. (2006). *Roc the mic right: The language of hip hop culture*. Routledge.

Althusser, L. (1971). Ideology and ideological state apparatuses (notes towards an investigation) (B. Brewster, Trans.). In, *Lenin and philosophy and other essays* (pp. 79–87). Monthly Review Press.

Ames, M. G. (2018). Deconstructing the algorithmic sublime. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718779194

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

Aydin, C. & de Boer, B. (2020). Brain imaging technologies as source for extrospection: Self-formation through critical self-identification. *Phenomenology and the Cognitive Sciences*, *19*(4), 729–745. https://doi.org/10.1007/s11097-020-09667-1

Baria, A. T. & Cross, K. (2021). The brain is a computer is a brain: Neuroscience's internal debate and the social significance of the computational metaphor. *CoRR*, *abs/2107.14042*. https://arxiv.org/abs/2107.14042

Beerends, S. & Aydin, C. (2021). Negotiating authenticity in technological environments. *Philosophy & Technology*, *34*(4), 1665–1685. https://doi.org/10.1007/s13347-021-00480-5

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*, 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*, 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.

Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547–591. https://doi.org/10.1007/s11229-008-9375-x

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

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bishop, S. (2018). Anxiety, panic and self-optimization: Inequalities and the youtube algorithm. *Convergence*, *24*(1), 69–84. https://doi.org/10.1177/1354856517736978

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

Blake, R. (2016). Toward heterogeneity: A sociolinguistic perspective on the classification of black people in the twenty-first century. In S. H. Alim, J. R. Rickford & A. F. Ball (Eds.), *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190625696.003.0009

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

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., ... et al. (2021). On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258

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

Boumans, M. & Leonelli, S. (2020). From dirty data to tidy facts: Clustering practices in plant phenomics and business cycle analysis. In S. Leonelli & N. Tempini (Eds.). Springer. https://doi.org/10.1007/978-3-030-37177-7_5

Boutonnet, B., Dering, B., Viñas-Guasch, N. & Thierry, G. (2013). Seeing objects through the language glass. *Journal of Cognitive Neuroscience*, *25*(10), 1702–1710. https://doi.org/10.1162/jocn_a_00415

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

Burrows, R. J. (2009). Urban informatics and social ontology. In M. Foth (Ed.), *Handbook of research on urban informatics: The practice and promise of the real-time city* (pp. 450–454). IGI Global. https://doi.org/10.4018/978-1-60566-152-0.ch030

Caplan, R. & boyd danah, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of facebook. *Big Data & Society*. https://doi.org/10.1177/2053951718757253

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

Clowes, R. W., Gärtner, K. & Hipólito, I. (2021). The mind technology problem and the deep history of mind design. In R. W. Clowes, K. Gärtner & I. Hipólito (Eds.), *The mind-technology problem: Investigating minds, selves and 21st century artefacts*. Springer, Cham. https://doi.org/10.1007/978-3-030-72644-7_1

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

Coeckelbergh, M. & Funk, M. (2018). Wittgenstein as a philosopher of technology: Tool use, forms of life, technique, and a transcendental argument. *Human Studies*, *41*, 165–191. https://doi.org/10.1007/s10746-017-9452-6

Cohen, S. M. & Reeve, C. D. C. (2021). Aristotle's metaphysics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/aristotle-metaphysics/

Cooperrider, K. & Núñez, R. (2016). How we make sense of time. *Scientific American Mind*, *27*(6), 38–43. https://doi.org/10.1038/scientificamericanmind1116-38

Crawford, K. (2021). *Atlas of ai: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, K. & Schultz, J. (2019). Ai systems as state actors. *Columbia Law Review*, *119*(7), 1941–1972. https://www.jstor.org/stable/26810855

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

D'Amour, A., Heller, K. A., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C. Y., Mincu, D., ... Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. https://arxiv.org/abs/2011.03395

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

Descartes, R. (2008). *Meditations on first philosophy: With selections from the objections and replies* (M. Moriarty, Trans.). Oxford University Press.

Di Paolo, E. A., Cuffari, E. C. & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. The MIT Press.

Dick, S. (2015). Of models and machines: Implementing bounded rationality. *Isis*, *106*(3), 623–634. https://doi.org/10.1086/683527

D'Ignazio, C. & Klein, L. F. (2020). *Data feminism*. MIT Press.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Eichhorn, K. (2019). *The end of forgetting: Growing up with social media*. Harvard University Press.

Elkins, K. & Chun, J. (2020). Can GPT-3 pass a writer's turing test? *Journal of Cultural Analytics*, *5*(2). https://doi.org/10.22148/001c.17212

Ensmenger, N. (2011). Is chess the drosophila of artificial intelligence? a social history of an algorithm. *Social Studies of Science*, *42*(1), 5–30. https://doi.org/10.1177/0306312711424596

Ensmenger, N. (2013). Computation, materiality, and the global environment. *IEEE Annals of the History of Computing*, *35*(3). https://doi.org/10.1109/MAHC.2013.33

Ensmenger, N. (2018). The environmental history of computing. *Technology and Culture*, *59*(4), 57–533. https://doi.org/10.1353/tech.2018.0148

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

Fedus, W., Zoph, B. & Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. https://arxiv.org/abs/2101.03961

Firth, R. (1971). *Elements of social organisation* (1st ed.). Routledge. https://doi.org/10.4324/9781315017525

Fischer, S. W. S. (2020). The difference between explainable ai and interpretable ai: Towards causal reasoning [unpublished manuscript].

Floridi, L. & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1

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.

Frické, M. (2015). Big data and its epistemology. *J Assn Inf Sci Tec*, *66*(4), 651–661. https://doi.org/10.1002/asi.23212

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)

Fuss, D. (1989). *Essentially speaking: Feminism, nature & difference*. Routledge.

Gallagher, S. (2011). *The oxford handbook of the self*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548019.001.0001

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

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*. https://dl.acm.org/doi/10.5555/2969033.2969125

Guo, L. N., Lee, M. S., Kassamali, B., Mita, C. & Nambudiri, V. E. (2021). Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-a scoping review. *Journal of the American Academy of Dermatology*. https://doi.org/10.1016/j.jaad.2021.06.884

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

Harvard, S. & Winsberg, E. (forthcoming). The epistemic risk in representation. *Kennedy Institute of Ethics Journal*.

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

Heinrich, J., Heine, S. J. & Norenzayan, A. (2010). The weirdest people in the world? *RatSWD Working Paper*, (139). https://doi.org/10.2139/ssrn.1601785

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

Hong, S.-H. (2021). Technofutures in stasis: Smart machines, ubiquitous computing, and the future that keeps coming back. *International Journal of Communication*, *15*, 1940–1960.

Husserl, E. (1970). *The crisis of european sciences and transcendental phenomenology: An introduction to phenomenological philosophy* (D. Carr, Trans.). Northwerstern University Press.

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*, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

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*, 306–316. https://doi.org/10.1145/3351095.3372829

Johnson, M. (2008). What makes a body? *Journal of Speculative Philosophy*, *22*(3), 159–169.

Jones, M. L. (2018). How we became instrumentalists (again). *Historical Studies in the Natural Sciences*, *48*(5), 673–684. https://doi.org/10.1525/hsns.2018.48.5.673

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, *11*(100). https://doi.org/10.1038/s41598-020-79310-1

Kowalski, R. (1979). Algorithm = logic + control. *Communications of the ACM*, *22*(7), 424–436. https://doi.org/10.1145/359131.359136

Kudina, O. (2022). Speak, memory: The postphenomenological analysis of memory-making in the age of algorithmically powered social networks. *Humanities and Social Sciences Communications*, *9*(7). https://doi.org/10.1057/s41599-021-00972-x

Lambrecht, A. & Tucker, C. (2019). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Lash, S. (2007). Power after hegemony: Cultural studies in mutation? *Theory, Culture & Society*, *24*(3), 55–78. https://doi.org/10.1177/0263276407075956

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

Leonelli, S. (2014). What difference does quantity make? on the epistemology of big data in biology. *Big Data & Society*, 1–11. https://doi.org/10.1177/2053951714534395

Luccioni, A. & Viviano, J. (2021). What's in the box? an analysis of undesirable content in the common crawl corpus. *Proceedings of the 59th An-*

nual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 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*, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

Malafouris, L. (2021). How does thinking relate to tool making? *Adaptive Behavior*, *29*(2), 107–121. https://doi.org/10.1177/1059712320950539

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*, 2145–2148. https://doi.org/10.1145/3340531.3412152

McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, *27*(4). https://doi.org/10.1609/aimag.v27i4.1904

McGuffie, K. & Newhouse, A. (2020). The radicalization risks of gpt-3 and advanced neural language models. https://arxiv.org/abs/2009.06807

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

Miceli, M., Posada, J. & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proc. ACM Hum.-Comput. Interact.*, *6*(GROUP), 1–14. https://doi.org/10.1145/3492853

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

Mikolov, T., Yih, W.-t. & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Compu-*

tational Linguistics: Human Language Technologies, 746–751. https://aclanthology.org/N13-1090

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*, 220–229. https://doi.org/10.1145/3287560.3287596

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

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Moeller, H.-G. & D'Ambrosio, P. J. (2021). *You and your profile: Identity after authenticity*. Columbia University Press.

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)*, 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

Ng, A. Y. & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 841–848. https://dl.acm.org/doi/10.5555/2980539.2980648

Nida-Rümelin, J. (2022). Digital humanism and the limits of artificial intelligence. In H. Werthner, E. Prem, E. A. Lee & C. Ghezzi (Eds.), *Perspectives on digital humanism*. Springer, Cham. https://doi.org/10.1007/978-3-030-86144-5_10

Nietzsche, F. (1967). *The will to power* (W. Kaufmann, Ed.; W. Kaufmann & R. J. Hollingdale, 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

Nilsson, N. J. (2009). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.

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

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)*, 2085–2094.

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

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training, 12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners, 24.

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)*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. & Sutskever, I. (2021). Zero-shot text-to-image generation. https://arxiv.org/abs/2102.12092

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. & Lee, H. (2016). Generative adversarial text to image synthesis. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (pp. 1060–1069). PMLR. https://proceedings.mlr.press/v48/reed16.html

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

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

Schramowski, P., Friedrich, F., Tauchmann, C. & Kersting, K. (2021). Interactively generating explanations for transformer language models. *CoRR*, *abs/2110.02058*. https://arxiv.org/abs/2110.02058

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 1–12. https://doi.org/10.1177/2053951717738104

Srinivasan, R. (2012). Re-thinking the cultural codes of new media: The question concerning ontology. *New Media & Society*, *15*(2). https://doi.org/10.1177/1461444812450686

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

Sutherland, T. (2014). The monument to a crisis: Nietzsche and the industrialization of creativity. *Third Text*, *28*(6). https://doi.org/10.1080/09528822.2014.970774

Sutterlüty, F. & Tisdall, E. K. M. (2019). Agency, autonomy and self-determination: Questioning key concepts of childhood studies. *Global Studies of Childhood*, *9*(3), 183–187. https://doi.org/10.1177/2043610619860992

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*(236), 433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

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

Weizenbaum, J. (1966). Eliza — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

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

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R. & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

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.

Zaidel, D. W. (2018). Culture and art: Importance of art practice, not aesthetics, to early human culture. In J. F. Christensen & A. Gomila

(Eds.), *The arts and the brain* (pp. 25–40). Elsevier. https://doi.org/10.1016/bs.pbr.2018.03.001

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00382-7

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power*. Profile Books.

Online Resources

Brockman, G., Murati, M., Welinder, P. & OpenAI. (2020, June 11). *Openai api*. Retrieved August 18, 2021, from https://openai.com/blog/openai-api/

Crawford, K. & Joler, V. (2018, September 7). *Anatomy of an ai system: The amazon echo as an anatomical map of human labor, data and planetary resources*. AI Now Institute and Share Lab. https://anatomyof.ai

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

Goode, L. (2021, April 6). *I called off my wedding. the internet will never forget*. Retrieved June 2, 2021, from https://www.wired.com/story/weddings-social-media-apps-photos-memories-miscarriage-problem/

GPT-3. (2020, September 8). *A robot wrote this entire article. are you scared yet, human?* Retrieved October 27, 2021, from https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Langston, J. (2021, May 21). *From conversation to code: Microsoft introduces its first product features powered by gpt-3*. Retrieved October 28, 2021, from https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/

Lowe, R. & Leike, J. (2022, January 27). *Aligning language models to follow instructions*. Retrieved February 25, 2022, from https://openai.com/blog/instruction-following/

Marcus, G. & Davis, E. (2020, August 22). *GPT-3, bloviator: Openai's language generator has no idea what it's talking about*. Retrieved September 22, 2021, from https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

OpenAI. (2021, November 18). *Openai's api now available with no waitlist*. Retrieved January 14, 2022, from https://openai.com/blog/api-no-waitlist/

PRIMO.ai. (2020, October 11). *Discriminative vs. generative*. Retrieved February 21, 2022, from https://primo.ai/index.php?title=Discriminative_vs._Generative

Radford, A. (2018, June 11). *Improving language understanding with unsupervised learning*. OpenAI. Retrieved December 13, 2021, from https://openai.com/blog/language-unsupervised/

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M. & Sutskever, I. (2019, February 14). *Better language models and their implications*. OpenAI. Retrieved December 13, 2021, from https://openai.com/blog/better-language-models/

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/

Scott, K. (2020, September 22). *Microsoft teams up with openai to exclusively license gpt-3 language model*. Retrieved October 28, 2021, from https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/

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/

Weng, L. (2018, August 12). *From autoencoder to beta-vae*. Retrieved February 26, 2022, from https://lilianweng.github.io/posts/2018-08-12-vae/