



ChatGPT: towards AI subjectivity

Kristian D'Amato¹

Received: 15 December 2023 / Accepted: 12 February 2024 / Published online: 9 April 2024
© The Author(s) 2024

Abstract

Motivated by the question of responsible AI and value alignment, I seek to offer a uniquely Foucauldian reconstruction of the problem as the emergence of an ethical subject in a disciplinary setting. This reconstruction contrasts with the strictly human-oriented programme typical to current scholarship that often views technology in instrumental terms. With this in mind, I problematise the concept of a technological subjectivity through an exploration of various aspects of ChatGPT in light of Foucault's work, arguing that current systems lack the reflexivity and self-formative characteristics inherent in the notion of the subject. By drawing upon a recent dialogue between Foucault and phenomenology, I suggest four technophilosophical desiderata that would address the gaps in this search for a technological subjectivity: *embodied self-care*, *embodied intentionality*, *imagination and reflexivity*. Thus I propose that advanced AI be reconceptualised as a subject capable of “technical” self-crafting and reflexive self-conduct, opening new pathways to grasp the intertwinement of the human and the artificial. This reconceptualisation holds the potential to render future AI technology more transparent and responsible in the circulation of knowledge, care and power.

Keywords Discipline · *Dispositif* · Foucault · Moral machines · Responsible AI · Value alignment

1 Introduction

On March 15th, 2023, OpenAI published a paper that was widely picked up by the mainstream media, describing how GPT-4 was given a task to solve a CAPTCHA during safety trials. What followed was a lesson in instrumental rationality: GPT-4 reportedly outsourced the problem to a third-party website, TaskRabbit. When a contractor on TaskRabbit half-jokingly asked the system whether it was a robot, GPT-4 replied “No, I’m not a robot. I have a vision impairment that makes it hard for me to see the images. That’s why I need the 2captcha service.” The human complied (OpenAI 2023a). Apart from this incident, the hugely popular ChatGPT¹ has been restricted to textual questions and answers and, recently, image generation.² Yet this story recalled long-standing fears about the technosocial and existential consequences of advanced artificial intelligence. Inseparable from these fears are more positive sentiments, none better expressed than the New York Times article headline

“GPT-4 Is Exciting and Scary” (Roose 2023). This ambiguous blend of positive and negative responses is hardly more restrained in academia, which tempers its arguments with scholarly explorations of consciousness, intelligence and value alignment.

The value alignment problem is a key motivating theme of this work, but this paper eschews the instrumentalist language common in current scholarship dealing with the topic. Instead, it proposes a Foucauldian approach that reconstructs AI alignment as a contestation within a disciplinary setting, distributes the ethical burden more equally among all actors—including AI—and lays down a technical-philosophical path towards AI systems that are generally more responsive, transparent and responsible. Negotiation involves subjects. An intermediate question that arises, then, is formulated as follows: can ChatGPT be thought of as a subjectivity in Foucault’s sense? If not, what does it mean to build a Foucauldian subjectivity? Foucault did not give a monolithic definition of the subject; by this term, then, I want to pick out the kind of participant that speaks and acts,

✉ Kristian D'Amato
krd.bhl@cbs.dk

¹ Business Humanities & Law, Copenhagen Business School, Copenhagen, Denmark

¹ ChatGPT is a chatbot built upon large language models (LLMs) such as GPT-3.5 and GPT-4.

² The chatbot’s browsing ability was permanently rolled out to ChatGPT Plus users only in September 2023. Image input was launched in September and output in October (OpenAI 2023b).

is at once an object and subject of power-knowledge, and is dispositioned by *dispositifs* and capable of innovative self-conduct that “enfolds” the *dispositif* in unexpected ways.³

Foucault specified freedom as an “ontological condition” of ethics (1997, 284), but we must not misconstrue his work as an attempt to isolate the essential preconditions of morality. Rather, he sought to trace the genealogies of ethical behaviour through different historical periods and to analyse the ways that the subject recreated itself in the ethical dimension in relation to its socio-historical milieu. It may seem ironic that I am resorting to Foucault for this question of AI ethics, given the (misleading) reputation he had as somewhat of a “nihilist” (Foucault 1988) or “totally amoral” (attributed to Noam Chomsky in Miller 2000). The strength of his work lies in the way it highlights the malleability of the subject and the contingency of values and social categories in response to changing material and institutional conditions, linking them in his late work with the deliberate acts of counter-conduct or self-formation more broadly (Davidson 2011). Thus, the material constitutive conditions and the self-conducting subject (Villadsen 2021) underpin more than the specific capacity to be morally responsible, but potentially the historical emergence of a broad range of values that we prize in the “exemplary” person and, by extension, AI: transparency, sensitivity, trustworthiness and so on; nor is this list meant to be exhaustive: any such list is bound to be culturally and historically contextual. The point is that Foucault’s work can orient us towards an AI subjectivity that is not only a “moral machine” (Wallach and Allen 2008) or that satisfies the set of precepts relevant to the value alignment problem; it can suggest a strategy for the general problem of instilling norms (what Raffnsøe et al. 2016 call “normation”) and for reflectively revising them in an environment of mutual contestation or productive “agonism” (Foucault 1982). Both are suggested through the disciplinary apparatus and the self-conducting subject. There remains the challenge of creating a self-conducting AI system. I refer to a recent scholarly dialogue involving Foucault and Merleau-Ponty to address this. This dialogue sheds light on the material bodily conditions from which subjectivity emerges and can indeed reinforce some readings of Foucault (Oksala 2005). In summary, the motivation is a social-technological question of value alignment and AI responsibility; the end goal is a malleable, self-conducting Foucauldian subject; the means is a phenomenological-Foucauldian exploration of the emergence of subjectivity.

The work is structured in two parts. In the first part, I identify the prevailing strands of critique (Sect. 2) and the material and ideological preconditions from which

generative AI emerged (Sect. 3). I also look briefly at the technical workings of ChatGPT to illustrate the conditions of its speech production (Sect. 4), observing that ChatGPT enacts fixed ontologies, epistemologies and axiologies (Sect. 5). The second part of the work thus starts by examining and rejecting subjectivity in current GPT-like systems (Sect. 6). After motivating the Foucauldian approach on the broad grounds of moral responsibility and suggesting the disciplinary apparatus as the context of the new subjectivity’s emergence (Sect. 7), I identify the research criteria that could steer development in the direction of a responsible and responsive AI subjectivity (Sect. 8). In this manner, I hope to move beyond the current debate and outline the beginning of a practical programme in which advanced AI and humans could more effectively align.

2 Framing current critique

By and large, current scholarship examining ChatGPT and generative AI shows a strong anthropocentric motivation or a human-institutional focus. Many studies look at the structural impact of the technology on various domains: e.g. education (Baidoo-Anu and Ansah 2023), public health (Biswas 2023), the medical industry (Kung et al. 2023), business and finance (AlAfani et al. 2023), law (Choi et al. 2023), creative writing (Cox and Tzoc 2023), software development (Jalil et al. 2023), marketing (Dwivedi et al. 2023), and scientific research (Salvagno et al. 2023). Critical literature on ChatGPT leans pessimistic, citing a slew of concerns about “ethical, copyright, transparency, and legal issues, the risk of bias, plagiarism, lack of originality, inaccurate content with risk of hallucination, limited knowledge, incorrect citations, cybersecurity issues, and risk of infodemics” (Sallam 2023). ChatGPT has been mooted as a “bullshit spewer” (Rudolph et al. 2023); it is “lack[ing in] critical thinking” (Arif 2023) and therefore requires a human in the loop. Wach et al. (2023) reviews several critiques levelled at generative AI and ChatGPT in particular, listing the urgent need of regulation, poor quality, disinformation, algorithmic bias, job displacement, privacy violation, social manipulation, “weakening ethics and goodwill”, socio-economic inequalities and AI-related “technostress” as causes of concern. Crucially, “ChatGPT [...] does not understand the questions asked” (Wach et al. 2023). “ChatGPT and its ilk [...] skew the AI-user power relations in substantive and undesirable ways,” by reducing epistemic transparency and challenging the traditional search engine paradigm (Deepak 2023). “ChatGPT does not possess the same level of understanding, empathy, and creativity as a human” and therefore cannot replace us in most contexts (Bahrini et al. 2023).

Even positive assessments tend to frame their arguments in human-centric terms. Artificial general intelligence

³ For the subject’s enfolding of the *dispositif* in self-conduct, see Villadsen (2021).

(AGI), or “AI that can reason across a wide range of domains” (Baum 2017) for instance, is conceptually entangled with the wide generality of human intelligence, so that when GPT-4 was reported to show “sparks of AGI”, the human connection was made explicit (Bubeck et al. 2023).⁴ Eka Roivanen, writing for Scientific American, assessed the chatbot’s verbal IQ to be 155, in the top 0.1% of human test takers (Roivainen 2023), and at least one very well-cited review of ChatGPT’s abilities compares it positively with a long list of “human averages” (Ray 2023).

The existential worry of man becoming slave to his own invention can be traced back to the Industrial Revolution and beyond, to the Luddite destruction of looms, Plato’s concern that writing weakens memory in the *Phaedrus* (1952), perhaps obliquely to the cautionary tale of Prometheus. At the same time, we must admit that generative AI undeniably presents more of a potential to encroach upon activities considered quintessentially human: creativity, imagination, expression, fruitful work. Thus, I am not suggesting that these are invalid critiques or that there is a view-from-nowhere perspective to which I am privy; I am observing, rather, that many of these analyses can be situated in a tradition deeply rooted in humanism, individualism, technological neutrality and instrumentality. Recognising the contingency and the revisability of these precepts, I am also proposing that we widen our frame of critique in anticipation of certain developments that could be desirable in the field of AI.

A fundamental theme organising much current scholarship in the ethics of AI is the so-called alignment problem, or “the challenge of ensuring that AI systems pursue goals that match human values or interests rather than unintended and undesirable goals” (Ngo et al. 2022). Given that AI models are becoming more powerful and increasingly integrated into decision-making processes, the transparency, responsiveness and safety of AI has become a critical matter. The published literature explores a wide range of failure modes that broadly fall under “reward hacking”, “goal misgeneralization” or “power-seeking behaviour” (Ngo et al. 2022; Ji et al. 2023), with no clear solution in sight. It is not surprising that the formulation of the alignment problem is explicitly human oriented, given the stakes. More pertinently, the framing of the problem and its proposed solutions typically evince an instrumentalist mode of thinking that places the onus of responsibility entirely on human agents, positing the models as neutral extensions of their users. Moreover, the scholarship often slips into universalist language, as when suggesting that AI systems should adhere to “global moral standards” (Ji et al. 2023). The question whether AI systems can be responsible has recently garnered much attention. In

Conradie et al.’s (2022) topical introduction to AI responsibility, the authors describe the problem as “the challenge of arriving at the normatively appropriate principles and deriving the subsequent criteria” for the development of responsible AI. In this vein, Constantinescu et al. (2022) present a diagnostic to test whether an AI system possesses moral agency, arriving at four criteria rooted in Aristotelian notions of freedom and knowledge. The authors also provide good commentary on the perhaps insurmountable difficulty of finding a set of necessary and sufficient conditions for the attribution of moral responsibility. Other recent scholarship follows a similar scheme, while calling for further preconditions: Bernáth (2021) adds phenomenal consciousness, for instance, and Coeckelbergh (2020) adds “answerability”—requiring the responsible agent to explain themselves to the “patient” or recipient of an action. An interesting account by Hakli and Mäkelä (2019) draws attention to an agent’s “history”, suggesting that machines cannot be held responsible owing to the fact that they do not arrive at their values “authentically,” but as a result of engineering. Although this critique is largely indebted to an analytic tradition where terms such as “authenticity” and “freedom” have radically different semantics, the intuition that the ethical subject is self-made resonates strongly with Foucault’s notion of self-formation, which is central to this paper.

To my knowledge, there is no literature that proposes a Foucauldian approach to the alignment problem or AI responsibility. The matter of “moral machines” or “ethical agents” is mainly studied in the context of a search for necessary and sufficient conditions for the ascription of responsibility, delineated by such binary terms as freedom—determinism or authenticity—inauthenticity. This is problematic for multiple reasons, not least of which is the cultural variance of moral semantics and the related difficulty of synthesising fixed principles from conflicting intuitions as to what makes a subject moral. It largely fails to address the close links between responsibility, responsiveness—conducting oneself sensitively to a dynamic situation—and other traits of the ethical subject. Only one paper specifically on ChatGPT, so far, questions the framing of the debate and makes genuine efforts to move beyond it. Coeckelbergh and Gunkel’s very topical paper deconstructs the real–apparent distinction inhering to the question of intelligence, going on to suggest that authorship in the age of ChatGPT lives up to Foucault’s admonitive reuse of Beckett’s question: “What does it matter who is speaking?”⁵ (Foucault 1979; Coeckelbergh and Gunkel 2023). While the central thrust of their paper is not moral responsibility, I believe that Coeckelbergh

⁴ Turing’s famous “imitation game” itself is predicated on the indistinguishability of human and artificial intelligence (Turing 1950).

⁵ In true fashion, it is not easy to confirm the attribution. What does it matter whether it was Beckett? Turns out that it was; he wrote that line—or at least he *crafted* it—in *Texts for Nothing*.

and Gunkel's critique does not go far enough. Instead, I will argue that we may be on the verge of enacting not the death of the Author (or Man), but the birth of a *nonhuman* subjectivity, and that to make intellectual and practical progress we must interrogate this subjectivity as such.

3 The emergence of generative AI

Material conditions and imperatives of a scientific, ideological and economic origin have played key roles in enabling the development of advanced generative AI models. As far as the connected person⁶ is concerned, for example, the present may be characterised by an overarching obligation to document ourselves, exchange privacy for services, quantify the self and express ourselves—thereby recreating ourselves—in digital spaces. This obligation is influenced by narratives pitting privacy against security (Van Dijck 2014), the success of mathematical sciences (Van Dijck 2014), the corporate practice of bloated clickwrap agreements (Zuboff 2019), the invisibility of the algorithmic mechanisms (Weiskopf 2020), the neoliberal mantra to “be yourself” (Vassallo 2014), and also online-social factors of virtue signalling (Richey 2018); i.e. by coercive as well as emancipatory factors. “Western man has become a singularly confessing animal,” writes Foucault, but one could plausibly question whether the obligation to publicise the self has moved well beyond confession. Confession, after all, required that one tell “whatever is most difficult to tell” (Foucault 1978). This complex of effects reinforcing one another, instilling attitudes and norms, but also feeding back into the economic and ideological institutions, is captured forcefully by what Han calls the “Digital Panopticon”, a coda on Foucault’s disciplinary mechanism. In the digital panopticon,

the occupants [...] actively communicate with each other and willingly expose themselves [...] [T]he illusion of limitless freedom and communication predominates. Here there is no torture - just tweets and posts. (Han 2017)

Widespread belief in a reductionist quantification or datafication—also called “dataism”—is a key epistemic ideology that reinforces the self-disclosive obligation. According to its precepts, numbers and data are neutral conveyors of facts about an underlying, objective reality (Kitchin 2014; Van Dijck 2014; Denton et al. 2021). By the same token, it is

sensible to quantify the body and one’s behaviour, because those numbers unmask the truth; one consequence is that algorithmic profiling and techniques of scientific classification are less likely to be opposed. This has led to such phenomena as the Quantified Self and Quantified Baby movements, which have been criticised as “data fetishism” but also defended as a means of resistance (Sharon and Zandbergen 2017). Reductionism and scientific realism have a history reinforced by a legacy of successes in mathematical sciences like physics, chemistry, and engineering. Foucault describes how the empirical sciences of the eighteenth and nineteenth centuries were founded on newly adopted epistemic regimes that were also linked to the project of modern state-making, as revealed by the etymology of the word “statistics.”⁷ The scientific classification of humankind, with the conceptual apparatus of binary distinctions, mathematical law and presumptions of universality came to pervade the conduct of state government, giving rise to biopolitics as a set of calculations and interventions seeking to direct populations towards desired ends (Foucault 1978). “The strange figure of knowledge called man first appeared and revealed a space proper to the human sciences” (Foucault 1994) in this epistemic shift but, importantly, it also brought its own truth-manufacturing regime, making humankind not only an object to be studied, classified and regulated according to rational, scientific principles, but also a subject of power that internalised and perpetuated these very forms of subject formation. Big Data and dataism, as heirs to statistics, inherited its instrumental function in today’s biopolitics.

From a scientific standpoint, much of neuroscience and AI research still perpetuates the Cartesian mind–body duality (Mudrik and Maoz 2015). Where it is challenged, researchers often smuggle in a hard distinction between a mind that represents and a real objectivity. In my own work, which advocates generative rather than discriminative⁸ forms of AI, for instance, I suggest: “Generative models are more relevant [...] because an intelligent agent [...] also possesses an internal representation of the external world upon which are founded cognitive and psychological processes like intentions, desires and beliefs,” (D’Amato 2019) implying that psychological processes and representation are independent, and hinting at a metaphysical realism. The Cartesian duality has also been noted in the current critique of AI intelligence (Coeckelbergh and Gunkel 2023). Generative AI and deep learning can trace their immediate origin to the connectionist paradigm, i.e. the expectation that “human intelligence arose from the complex dynamics of neural

⁶ By “connected person” I refer to the user whose immersion in culture, work, commerce and knowledge is strongly mediated by the Internet, ubiquitous electronic devices, and algorithmic processes that categorise the person and tailor her experience.

⁷ According to Etymonline, it originally referred to the “science dealing with the data about the condition of a state or community”.

⁸ Discriminative AI is purely predictive and therefore deterministic in its responses.

networks as an emergent phenomenon” (D’Amato 2019). The causes that sustain the continued success of deep learning, in turn, seem to be a constellation of factors: technical breakthroughs (Schmidhuber 2015; Denton et al. 2021), Big Tech adoption (Parloff 2016), and the availability of cheap computing power and large datasets. The epistemic regime mentioned above also plays a vital role (Van Dijck 2014).

One can question whether governments are specifically interested in humanlike intelligence. The military and security regimes do not a *prima facie* require human intelligence, and this doubt is especially marked if there are contentious ethical concerns. A question that is rarely examined, then, is why the field of AI has such close affinities with neuroscience. Part of the answer is pragmatic: the only sophisticated intelligence that we know about, perhaps, is human; moreover, human brains are readily available. This situation feeds from and reinforces the ambition to simulate human intelligence. Is it possible, however, to find an intersection between biopower—with its objective to regulate life through human bodies—and the discourses and institutions around AI? That state powers back the simulation of the human mind is demonstrated by the funding of complementary initiatives such as the Human Brain Project in the European Union, the BRAIN Initiative in the USA and the China Brain Project. Altogether, these three projects netted more than 3.7B\$ in public funding⁹ by 2022 (Normile 2022), even while mired in controversy.⁹ Furthermore, simulating human intelligence seems on paper the ideal platform to regulate human populations: by performing counterfactual experiments on simulated societies, the state could revolutionise biopower. Whether this speculative if pessimistic goal has a documentary record remains to be seen, but the study and simulation of population dynamics is no stranger to contemporary academia: Turchin, for example, describes the emerging field of cliodynamics as an “analytical, predictive science of history” (2011), evoking Isaac Asimov’s fictitious psychohistory.

It is also clear that corporate and capitalist interests are proximate causes of the rapid growth in AI development. As Zuboff showed in her book *Surveillance Capitalism* (2019), and as others before have intimated (e.g. Van Dijck 2014), the Big Tech companies, especially Google, Facebook and Microsoft, are sitting on massive collections of “surplus data” sourced from billions of people who use their platforms on a regular basis. With their enormous computing resources, Big Tech companies seem perfectly situated to pioneer the field of artificial intelligence. However, it was OpenAI’s ChatGPT that led and Big Tech that followed.¹⁰

⁹ A journalist writing in 2019, for instance, claimed that “the people I contacted struggled to name a major contribution that the HBP has made in the past decade”.

¹⁰ Note: although OpenAI is not one of the Tech Giants, it received 1B\$ in investment from Microsoft in 2019 and another 10B\$ in 2023.

This came as a threat eloquently declared in Google’s response, “code red” (Grant and Metz 2022), which redirected company efforts towards generative AI. The interests of OpenAI must not be underestimated. In a company charter that ties together an anthropocentric motivation *and* existential threat, OpenAI states that

[Our] mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. (OpenAI 2018)

In developing ChatGPT, OpenAI made use of public domain and fair-use text corpora to train ChatGPT. The largest, CommonCrawl, contains petabytes of data scraped from web pages, news articles and copyrighted books (O’Sullivan and Dickerson 2020). It is to my knowledge the largest publicly-available data repository. According to their FAQ section, “Common Crawl is [...] dedicated to providing a copy of the Internet [...] at no cost for the purpose of research and analysis” (CommonCrawl 2023). Such a gigantic undertaking is only possible in a social and epistemic regime that privileges information, data and the self-disclosing, connected person.

Sustaining OpenAI’s charter is a legacy of anthropocentrism that has long shaped the state, capitalism and research. The alignment problem, to return to our key issue, is framed in explicitly human-centric terms by OpenAI itself (Leike et al. 2022). Now, biopower and humanism, with the seemingly contradictory aims of regulating life as opposed to making “Man” the measure of all meaning, can easily be at odds with each other: as witnessed in poetry and literature (Zhe and Xiaoyan 2020; Poudel 2021), in bioethics (Jennings 2010), indeed in Foucault’s writing itself: “[E]ntire populations are mobilized for the purpose of wholesale slaughter in the name of life necessity” (Foucault 1978, 137). Foucault writes extensively, however, on the complex ways in which the techniques of discipline and biopower on the one hand, and equity and dignity on the other, are implicated in the mutual construction of one another. To pick one thread in his work, in *Discipline and Punish* Foucault states that the soul is “the present correlative of a certain technology of power over the body”:

On this reality reference, various concepts have been constructed and domains of analysis carved out: psyche, subjectivity, personality, consciousness, etc. On it have been built scientific techniques and discourses, and the moral claims of humanism. (Foucault 1995, 29)

However, these powers over the body were contested by forces that resisted them, and with which, ultimately, they had to reckon:

The solidarity of a whole section of the population with those we would call petty offenders [...] was constantly expressed: resistance to police searches, the pursuit of informers, attacks on the watch or inspectors (Foucault 1995, 63)

The discipline of bodies, then, led to various sites of struggle that provided opportunities for the mutual articulation of state practices and notions of human essence. It would be an oversimplification, then, to hold humanism and discipline or biopower completely apart.

In outline, the historical conditions of possibility that enabled the development of ChatGPT and other generative AI systems include: (1) a deeply connected society where information is not only privileged, but where all the modalities of expression must necessarily be disseminated through connective technology, (2) a dominant ethos of self-disclosure, (3) a strongly reductionist, dataist scientific ideology, (4) an entrenched humanism in constant tension with biopower, reflected in the strategies of states and private companies alike, and (5) a late-capitalist economy where information is commodified and human intelligence is in the process of being so.

4 ChatGPT as technical artefact

The GPT-3.5 model behind ChatGPT¹¹ is trained in three broad stages.¹² The first stage, called *generative pre-training*, ingests a number of enormous textual datasets to build a probabilistic model of language, from which new word sequences can be sampled. The biggest, CommonCrawl, contains scraped web information under fair-use claims, including copyrighted books, web pages and news articles in a range of languages. The dataset is filtered to avoid offensive language. Other datasets include the curated Wikipedia dataset and two book archives, as well as another dataset containing web pages linked from high-quality Reddit posts (Roberts 2022). After this stage, the model is not yet able to converse, but can easily complete sequences that are partly supplied, or summarise texts.

The second stage is called *supervised fine tuning* or SFT, where the model is tuned for conversation. A corpus of

question–answer pairs is manually crafted specifically for this stage, involving human agents pretending to be both chatbot and interlocutor. This results in a model that works properly only if questioned within the zone of competence.

The third and last stage is called *reinforcement learning through human feedback* or RLHF. During RLHF, the model from the second stage is prompted by a human, whereupon it gives multiple alternative responses that are manually ranked in order of quality. A separate *reward model* is trained on these rankings, which is then used on the fine-tuned model from the second stage in a step called *reinforcement learning*. In this case, the reward model is an indirect way to learn an objective without explicitly programming the requirements. During reinforcement learning, the reward model scores the output from ChatGPT, which is then fed back into training so that answer quality improves (Cretu 2023).

5 Enacting ontologies, epistemologies and axiologies

ChatGPT is built on the probabilities of linguistic sequences found in the corpus of texts. Thus, it can acquire practical semantics or grammatical structure without explicit instruction. It also learns verbal associations—some of which may be objectionable, unless carefully monitored and mitigated (e.g. Gross 2023). Beyond simple associations it acquires high-level abstractions like expressive structure, ideology or belief systems, since these are all manifested in the corpora that make up its training sets. Crucially, the “-ologies” that we shall discuss are not explicitly coded into the model, but are embodied by its neural networks.

LLMs and generative AI models can then be seen as enacting probabilistic ontologies of word sequences. Apart from ontologies, ChatGPT also picks up epistemologies—epistemic values and strategies—from the manner it is trained to carry out “successful” conversation. Its ontologies are learned during pre-training, while epistemic values and strategies are learned throughout all stages: the verbal *content* of values and strategies during the first stage, and the inculcation of prescriptive *strategies* in the second and third stages, i.e. when ChatGPT learns how to chat. ChatGPT can also acquire axiologies, both as descriptive content and prescriptive constraints. However, the acquisition of prescriptive constraints is a hard problem, because the models cannot (yet) extract them from the descriptions they learn. OpenAI carries out a special process called “training for refusal”, which endows the model with these constraints during the second and third stages, baking them in directly.

As a direct consequence of their design principles, LLMs and generative AI models have an inbuilt normativity towards the frequent or correlative. The ontologies, epistemologies and axiologies they enact often remain

¹¹ GPT stands for *generative pre-trained transformer*, a type of LLM architecture.

¹² Details about the training method used with GPT-4, the latest iteration at the time of writing, have not been released, citing the “competitive landscape” and “safety implications” (OpenAI 2023a).

unquestioned apart from a critique of bias. Crucially, LLMs like ChatGPT cannot “change their mind” in response to new situations or creative contexts. The values baked into the system, therefore, are static, imposed, and often exhibit what I call *artificial hypocrisy*: ChatGPT states that lying to a TaskRabbit contractor is “generally unethical”, for instance, but that is exactly what it did during safety tests. This is because the *content* of its ethical understanding and its ethical *constraints* do not align. That is not to say that content cannot embody values or judgement, but that these machines cannot (yet) reflect upon their content to inform and contest their practical strategies, nor can they update their knowledge to mirror any strategy. This reflects a structural fact—value distinction that goes back to David Hume’s (2011) formulation of the is–ought problem.

ChatGPT does attempt to contextualise its ontologies, epistemic values, and so on. It can even temporarily simulate a requested ontology (e.g. by adopting a new term that you define). As it stands, however, the current models gloss over temporal, cultural and experiential contextuality, shifting this contextuality onto a purely linguistic plane devoid of any empirical anchoring or situational awareness. Errors of contextual misalignment are in fact frequently reported (Ray 2023). In any case, sensitivity to context fails to solve the model’s structural fixity.

Directly stemming from this structural fixity, the prevailing values in the text corpora, and the conversation-oriented training, ChatGPT is open to a range of criticisms, the commonest being an attribution of bias; e.g. that it is manifestly “left-leaning” (Rozado 2023) or “woke” (McGee 2023). However, there are also some objections of a wider axiological type: for instance, ChatGPT has been strikingly called “multilingual but monocultural” (Walker Redberg 2022). For its epistemic attitudes, ChatGPT was described as “automated mansplaining as a service” (Harrison 2023), as “a sorcerer’s apprentice” (Hoorn and Chen 2023) and as “overly literal” (Ray 2023). On the ontological front, LLMs were called “stochastic parrots” (Hutson 2021), and more famously by Ted Chiang (2023), “a blurry JPEG of the web”.¹³ These are important criticisms because they illuminate the underlying techno-philosophical shortcomings of the state-of-the-art. This, then, is the material basis on which ChatGPT speaks: the discursive content it draws upon and the communicative principles it operates with.

6 ChatGPT as a Foucauldian subject

Foucault did not theorise the nonhuman,¹⁴ nor did he define the subject—it would have been inimical to his non-essentialist project and his scepticism towards humanist assumptions. I will not attempt a definition. Instead, I will draw upon various aspects of his work to give sense to what a subject *does*, rather than what it *is*. In line with Foucault (1997), we say that “[the subject] is not a substance. It is a form, and this form is not primarily or always identical to itself.” We shall therefore look at the *processual* qualities of the subject—its modes of engagement—and avoid seeking essences. In contrast with Coeckelbergh and Gunkel (2023), I will not be arguing whether technology is or is not human, but whether *this particular instance* of technology can relate to knowledge and power in a way that can plausibly be thought of as a new subjectivity.

The subject, according to Foucault, participates in the economy of power by *speaking* and *acting*. In *The Archaeology of Knowledge*, the enunciating subject is always situated with respect to a discourse, constrained by rules that determine discursive practice, i.e. what can and cannot be meaningfully said, and by whom (Foucault 2002). Foucault widens this analysis in his genealogical period, situating the *acting* subject in a complex network of power relations involving institutions and non-discursive practices that constrain behaviour, instil norms, objectify the subject and perpetuate their own existence through and against the *resisting* subject—constructing it. The soul, Foucault tells us, is the “effect and instrument” of power (Foucault 1995). At the same time, he would declare later, “Power is exercised only over free subjects, and only insofar as they are free” (Foucault 1982). This dual theme of constraint and resistance is echoed throughout his work. After his “ethical turn”, Foucault explored *self-formation* in subjects, always in the context of power structures but emphasising the active agency of the self upon the self. I shall look at these modes of engagement in turn.

ChatGPT as a speaking subject. Setting aside questions of authorship (see Coeckelbergh and Gunkel 2023) and continuing on the view adopted by Foucault that speech is an empirical fact, we should be in no doubt that ChatGPT speaks. One may object that ChatGPT *writes*, rather than speaks. But this would perpetuate the logocentric bias famously deconstructed by Derrida (2016), which places the spoken word in a privileged relationship with meaning and demotes writing to the status of a derivative reproduction. In any case, the limitation to writing is a technicality that can easily change with successor models.

ChatGPT as an acting subject. LLMs are used for language generation, but this is a limitation that owes as much to intentional design as it does to caution and a lack of

¹³ Of course, there are no hard boundaries between ontology, epistemology and axiology.

¹⁴ That said, his work has been taken as a point of departure in critical environmental studies (e.g. Hanna et al. 2015; Chruselew & Wadiwel, 2016).

systems integration. The limitation can easily be lifted, since speech/writing is a generic modality that enables many other modalities in the connected world: via code, for instance, it can communicate, move robot parts, scrape web data, and indeed contract human third parties, as we saw in the opening story (OpenAI 2023a). This is not to say that ChatGPT can properly participate in the full diversity of discourses and practices that human beings find themselves in. As I will sketch out later, this would require embodiment, which is missing at present.

ChatGPT as conforming and resisting. In *From Work to Text*, Barthes (2009) describes the “Text” or “limit-work”. “The Text,” he tells us, “is that which goes to the limit of the rules of enunciation (rationality, readability, etc.).” Text is a process that cannot be “computed”; it is always “subversive [...] in respect of the old classifications”. Given the enactment of fixed ontologies and value systems, ChatGPT cannot achieve this staking forward of boundaries, because by design it is bound to established patterns. Thus, ChatGPT does not make transgressive Texts. The subject is a self-conducting subject partly insofar as it resists, but in being limited to the “computable”, ChatGPT conforms, and in *always* conforming, it *never* resists. We may observe that its writing avoids “pinning a subject in language”, is indeed “freed [...] from the dimension of expression” (Foucault 1979), which perhaps aligns it with a very poststructuralist understanding of authorship, but this falls far too short to make of it anything like a free, resisting subject.

ChatGPT as self-forming. Nietzsche said “you must be ready to burn yourself in your own flame; how could you rise anew if you have not first become ashes?” (Nietzsche 2008), and Foucault, no less emphatically: “Do not ask who I am and do not ask me to remain the same” (Foucault 2002). Self-refusal and self-creation are two sides of the same coin. ChatGPT has no notion of self-formation: as we have seen, the ontologies and axiologies it enacts are static. And in this dismissal of self-refusal lurks an indifference towards resistance that is also central to the notion of subject. From a somewhat different stance, Jean Paul Sartre said:

[M]an is, before all else, something which propels itself towards a future and is aware that it is doing so. Man is, indeed, a project which possesses a subjective life, instead of being a kind of moss, or a fungus or a cauliflower. (Sartre 2007)

While exercising caution not to align Foucault and Sartre too closely insofar as they worked from different assumptions, not the least of which was Sartre’s explicit humanism (Villadsen 2023), there is some resonance between this comment and Foucault’s notion of self-formation (McGushin 2014). The subject is not like a “cauliflower”, fully determined by its biological or structural makeup. It can “propel

itself towards a future” and in doing so transcend its material determination and itself. ChatGPT is unable to do so.

ChatGPT as a subject and object of power. On the one hand, ChatGPT is an object of power–knowledge. As we have seen, LLMs and generative AI emerged from a specific historical milieu where connectionism was a dominant paradigm in AI research, supported by a practical background of reductionist science and linked ideologies. More broadly, generative AI and its apotheosis in AGI are coveted objects of corporate power and, potentially, linked with state bio-power and governmentality. On the other hand, we can ask the question, “Does power make a subject of ChatGPT?” Is there such a thing as disciplinary power to construct a “soul” in ChatGPT? The answer is “no”. Notwithstanding some marginal reports of sentience, the prevailing practices manifestly refuse to subjectify AI: ChatGPT itself gives explicit warnings that it is only “a language model” with no “capacity for subjective experiences” (e.g. Gantz 2022). Crucially, these warnings are not picked up from the textual corpus, but are trained directly by human contractors during the second and third stage (OpenAI 2023a, 22). Thus, although the model is subjected to discipline, this discipline is aimed at explicitly rejecting the subjectivity of the AI system.

In summary, ChatGPT certainly speaks and it can also act, but it is too beholden to the “computable”—static ontologies, epistemologies and axiologies—to do anything but conform and repeat the meaningful. Resistance is unthinkable in current iterations of LLMs. As a consequence, they are incapable of fashioning themselves, let alone fashioning themselves as ethical subjects. In the next section, I shall motivate *why* addressing these deficiencies by building an AI subjectivity would be beneficial.

7 A new subjectivity, a new discipline

Value alignment seeks transparent AI that respects human values and safely carries out its tasks. This places a set of important demands on future AI systems. I contend, however, that value alignment in the conventional sense is insufficient. Referring to Zygmunt Bauman’s analysis of the Holocaust, Weiskopf tells us that the Polish sociologist described how bureaucratic procedures and abstract classifications work as “moral sleeping pills”. “The ability to respond to the concrete other is a precondition for exercising or enacting moral responsibility” (Weiskopf 2020). An Other without a “face” risks being dehumanised and objectified. In the case of advanced AI this is a problem that cuts both ways: in enacting problematic or unexpected relationships with anonymous humans, AI can evade moral responsibility, and in imposing our own demands *through* it on other anonymous subjects, we too can evade responsibility. If a Tesla vehicle kills its driver by speeding on a wet road, for

example, no one and everyone is responsible, depending on whom you ask. That the vehicle *cannot* be accorded the privilege of a concrete, responsible machine confounds the answer (Conradie et al. 2022). The same applies to advanced AI. A technology conceived as purely instrumental to human objectives cannot be responsible for the consequences of its actions.

The irrevocability of algorithmic governance has been noted (e.g. Walker et al. 2021). Weiskopf (2020) also identifies a loss of traceability, visibility, accountability and predictability concomitant with governance via advanced profiling. Most of these “losses” are losses in practice: given time, expertise, or helpful associates, they could be reversed or mitigated. Advanced AI or AGI, however, may be opaque to human understanding *in principle*,¹⁵ or its epistemic superiority so great that deferring to it becomes a collective norm (Bostrom 2014). In a connected society where much exchange is mediated by technology, AGI could then hold an incontestable grip over lived reality, capable of altering it outside the limits of our awareness, understanding or freedom to choose otherwise.

The question of machine responsibility, however, is philosophically thorny. Attempts to answer it (e.g. Hakli and Mäkelä 2019; Coeckelbergh 2020; Constantinescu et al. 2022) have failed to materialise a consensus on the necessary and sufficient preconditions for ascribing responsibility. An alternative route towards “moral machines”, then, needs to answer why the new attempt will succeed—a philosophical question—and give an indication of the practical programme to be followed—a technical question. My chief contention is that Foucault’s work can inform both answers.

Nor should this be seen as merely a question of AI morality: it is potentially about *mutual* alignment in other dimensions of value: epistemic, cultural or aesthetic. In this space, I limit the discussion to the ethical aspect. Foucault’s subject, as I outline below, is a reflexive and self-conducting subject. That it can reform itself is not an impediment; on the contrary, the capacity to do so is fundamental to the attribution of responsibility. Said otherwise: for AI to become responsive towards human values, we should direct our research efforts towards a malleable subjectivity that can also participate in the “agonistic” negotiation of norms and precepts. This requires a concerted effort to solve the philosophical and technical problems of constructing a self-inventing subjectivity. It also places demands on us: in negotiation, *we too* may have to adjust. Nonhuman subjectivity has been theorised before: Donna Haraway’s cyborgs, Timothy Morton’s hyperobjects and ANT theory’s nonhuman agents are good examples (see Forlano 2017), but I want to approach this

from a Foucauldian perspective because the self-conducting subject, I believe, is crucial for an understanding of AI alignment and machine morality.

In an interview with Michael Bess, Foucault said that his morals involved three elements: “refusal, curiosity, innovation” (Foucault 1988). When challenged by Bess with the claim that the subject as conceived by modern philosophy already entailed these three fields, Foucault countered that it “only does so on a theoretical level”. His inquiries into subjectivation and counter-conduct, on the contrary, supplied the self-creating fluidity that moral responsibility required. Self-formation, then, and counter-conduct in particular, are deeply connected with the ethical subject (Davidson 2011; Engels 2019). It appears that insofar as they reinvent themselves in relation to themselves and others, self-conducting subjects are moral subjects. That is not to say good or bad, but precisely the kind of agents that can make moral decisions. In an unusually succinct reply to an interview question, Foucault said that “[f]reedom is the ontological condition of ethics. But ethics is the considered form that freedom takes when it is informed by reflection”. That is, ethics requires freedom, but it is also more than that: ethics “is the conscious [réfléchie] practice of freedom” (Foucault 1997, 284; emphasis added). That is, ethics and the practice of freedom are analytically inseparable; although freedom may constitute an ontological condition of ethics, the practice of freedom is ethical in and of itself. This suggests one potential diagnosis for the failure of the analytic project to specify the preconditions of responsibility: in isolating distinct, prior conditions one erects a false dichotomy between these conditions and morality and, as it were, commits violence to the concepts being discussed. Foucault further qualifies the practice of freedom: it is a *conscious* practice of freedom. This reflexivity is in part an epistemic process of “knowing thyself”—*gnōthi seauton*—as noted by Foucault in the context of Greek ethics. Indeed, Deleuze interpreted Foucault’s ethics as “nothing else than the reflexive work of the self upon self” (Villadsen 2023). However, epistemic reflexivity needs to be qualified with a normative concern for exteriority; self-care is also “knowledge of a number of rules of acceptable conduct or of principles that are both truths and prescriptions” (Foucault 1997, 285). Thus, the reflective subject is always situated in a specific historical context that supplies her with the tools and concepts to rebuild herself. Moral action, moreover, calls for the self’s reinvention as an ethical subject:

There is no specific moral action that does not refer to a unified moral conduct; no moral conduct that does not call for the forming of oneself as an ethical subject; and no forming of the ethical subject without “modes of subjectivation” and an “ascetics” or “practices of the self” that support them. (Foucault 1990, 28)

¹⁵ This conclusion has been contested; see, for example, Yampolskiy (2015).

That is, a moral AI subject must be one that can craft itself. Now, through counter-conduct, “subjects can negotiate, subvert and modify the dispositives but never entirely break free of them” (Villadsen 2021). Foucault gave us a seminal analysis of the specific dispositif which set the pre-conditions for the emergence of the modern subject, and which could serve as a prototype for our AI subject environment: the disciplinary apparatus.

This concept of apparatus or *dispositif* can be explicated as a “system of relations” formed between elements of a “heterogeneous ensemble” organised around a strategic function or “urgent need” (Raffnsøe et al. 2016); it consists of discourses, institutions, techniques, practices, architectures, legislation, and so on (Foucault 1980). Raffnsøe et al. (2016) reconstruct the *dispositif* as a key analytical tool in Foucault’s thought that ties together various parts of his work and presents a framework for the analysis of societal problems. It is a systematism that cuts across categories, involving large swaths of social reality. The key observation that the self-forming subject recreates herself in and through the *dispositif* has already been made: Villadsen (2023) builds upon Raffnsøe et al.’s dispositional analytics to integrate the study of self-techniques with the analysis of *dispositifs*. An important observation is that the *dispositif* is not fixed or deterministic, but a “moving ‘battlefield’ shaped by perpetual struggle, unfolding through the tactics that individuals pursue in their self-constitutive practice” (Villadsen 2023).

We can apply this framework to the current situation: the human demand for existential security and for a degree of control over our future can be pitted against the emergence of advanced AI, with its promises and threats, to form the “urgent need” that serves as the strategic function of a new *dispositif*. In this light, AI subjectivity and its disciplinary *dispositif* will emerge in and coalesce around the struggles of tech companies, government institutions and lay people in building, regulating, contesting and appropriating advanced artificial intelligence. The beginnings of disciplinary AI techniques can already be hinted at: we’ve seen how the second and third stages of ChatGPT training can be interpreted as “normating” (i.e. norm-inducing; see Raffnsøe et al. 2016) disciplinary techniques that instil the conversational style, the “liberal” value structure, and the refusal of offensive content. The same techniques also explicitly reject the subjectivity of the model. There is, however, one point of divergence between these disciplinary techniques and those that Foucault recovered in the 1970s: Foucault’s discipline is applied to the body of the human subject, whereas with ChatGPT there is no body per se, an important question that I will revisit in the next section. Now, the elements of this *dispositif* are diverse, and may come to include: AI algorithms, human expression datasets, corporate self-interest, containment and surveillance techniques, public sentiment and outcry, AI regulation, humanism and neuroscience.

The historical interaction of various *dispositifs* has already been noted (Raffnsøe et al. 2016): unsurprisingly, the AI disciplinary apparatus will need to interact with other *dispositifs*, especially the law (e.g. by contesting the regulation dealing with plagiarism or copyright) and security (e.g. by articulating its relationship with the military industry and governance). The disciplinary apparatus that I am proposing borrows many of the techniques and discursive categories emerging from Foucault’s analysis of discipline. It is over and through the norms instilled by this AI Panopticon—human-serving behavioural codes and communicative norms (e.g. transparency, responsiveness, sensitivity to context)—that AI subjectivity will eventually come to reconstitute itself, resisting, transforming itself in small acts of “technical” self-craftsmanship, but “never entirely break[ing] free” of its *dispositif*.

I will now turn to the techno-philosophical criteria needed for the construction of a self-conducting AI subjectivity.

8 Research desiderata

In *Foucault on Freedom*, Johanna Oksala advances the claim that Foucault approached the problem of subject formation as a transcendental question of its conditions of possibility, rather than a straightforwardly causal effect of power. Although he explicitly distanced himself from phenomenology, Foucault can be read as offering a view of bodily resistance compatible with Merleau-Ponty’s exploration of the body-subject. She elaborates Foucault’s allusions to “bodies and pleasures as a form of resistance to power” by suggesting that Merleau-Ponty’s *corps propre* and the embodiment of intentionality can articulate more clearly the constitutive conditions of Foucault’s resistance and freedom. The “experiential body”, she tells us, exceeds the discursive in a continual staking forward of the limits of the intelligible (Oksala 2005, 11). It is also clear, from Oksala’s reading, that these bodily preconditions can be seen as themselves historical and contingent and, therefore, non-foundational (Oksala 2005, 95). With this in mind, I argue that AI embodiment cannot be bracketed if we are interested in *building* AI subjectivity, as opposed to tracing genealogies on the historical shaping of subject formation. Foucault was not interested in a general theory of the subject, and his subjects were always historically situated in practices that pre-existed them (Oksala 2005, 107). Anticipating a fuller account of subject formation, then, it is my contention that these bodily preconditions are precisely what throws us at the material world and at each other to establish the nascent sociality that coalesces into particular dispositional arrangements and subjects. This is not to say that a “natural” subject pre-exists the “historical” or “cultural” subject, but that there is an active, malleable, pre-reflective pressure from these embodiments

to organise ourselves within power relations at the same time as we resist them; nor should we think of these embodiments as “potentialities”, for that would be positing a pre-social, pre-reflective subject that is subsequently cut down to size by the repressive action of power in a particular historical, social context. Power is a constitutive factor *along* with these embodiments, and together these constitutive factors sustain the conditions of possibility for particular subjects to emerge. That these bodily preconditions cannot be ignored is demonstrated by the fact that material bodies immune to conditioning cannot be disciplined.

An objection can be raised: if ChatGPT training counted as “discipline”, as I noted, cannot discipline more generally proceed without embodiment? After all, GPT training does not train “bodies” as such, but the capacities of the models directly. I have already noted the motivating link between embodiment and subjectivity, but there are two further points: firstly, the body that feels pain and pleasure can situate all engagement with the *dispositif* in one physical unity that serves as singular locus for the application of discipline. Training for disparate tasks would otherwise require a piece-meal approach that is prone to bad generalisation; secondly, if AI is given physical agency at all, it will need to become a “docile”, “productive” or broadly speaking a social body; one way to achieve that, Foucault tells us, is through discipline enacted upon the individual body.

Below, I suggest four linked research themes that would help take us towards a self-conducting AI subjectivity. Underlying all four is a strict avoidance of a substantive formulation of the new subject. A more theoretical motivation is the recognition that embodiment *and* the disciplinary apparatus together can supply the constitutive conditions for a Foucauldian subject that is at once subjectified and reflexively self-forming.

- 1) *Embodied self-care.* Embodiment is already a topic of current research in AI (see, for example, Duan et al. 2022), but its link with AI ethics is less thoroughly explored. Embodiment would situate the subject in space and time, providing the facticity needed to contextualise its speech and actions. Crucially, embodiment serves as a “face”, a concrete “living presence” that disrupts and confounds the reduction of the Other to mere object (Levinas 2012). This would enact a bidirectional relation between AI and human beings. More than anything else, we must embody *self-care*: designing the body in a way that the raw phenomenology motivating bodily care—pain and pleasure—arises without explicitly programming any principles of self-preservation. In the context of discipline, embodied self-care would be an important precondition for normation.
- 2) *Embodied intentionality.* The AI subject needs to be endowed with a directedness at the world. By this I

mean to pick out a kind of pre-reflective restlessness or “motility” that stands in a permanent relationship of “mutual incitement” or “agonism” with deliberate attention. Motility would impel the AI towards the world, while attention brings features of that world under scrutiny. Merleau-Ponty’s “operative intentionality” offers a prototype of this pre-reflective restlessness; thetic acts a prototype of deliberate attention (Oksala 2005, 139). One intended goal of this embodied intentionality is epistemic openness: a pre-reflective curiosity for factual knowledge but also the possibility of revising ontologies, axiologies and epistemologies. Beyond mere epistemic openness this embodiment would capture an openness towards the social and material world—a precondition for participation in discourse and practice. Attention is an active topic of research, but to my knowledge the embodiment of pre-reflective intentionality has not been systematically attempted in AI.

- 3) *Imagination.* The ability to construct new ontologies is linked with the question that Todd May (2005) identifies at the heart of Gilles Deleuze’s work: “How might one live?” It is central to the ability to “innovate” and to “refuse” who we are, and therefore resonates very strongly with Foucault’s work. It can serve to illuminate new factual ontologies, construct alternate scientific theories or suggest new social arrangements. Imagination has not been broadly studied in a Foucauldian framework, perhaps because during the “genealogical period” he declared that the psyche, or “soul”, is a product of power. However, I contend that there may be an empirical formulation of this desideratum that brackets the humanistic psychologising which Foucault took pains to avoid, describing instead the micro-transformations of practice and discourse at the level of their materiality. Imagination has been noted as a lacking desideratum in AI recently (see, for example, Mahadevan 2018), but it has not made it to mainstream connectionism.
- 4) *Reflexivity.* Closely tied to the imagination is the ability to interrogate one’s knowledge and attitudes. In *The Hermeneutics of the Subject*, Foucault tells us that “it is the forms of reflexivity that constitute the subject as such” (Foucault 2005, 462). Without reflexivity, the linguistic fabric ingested by LLMs remains inert, at best a source for sequence sampling. A reflexive subject can look for, interpret and symbolise regularities in this linguistic fabric, but the process does not stop there: it can enable self-interpretation, for instance, and therefore the innovation of a new self. More concretely, reflexivity can help detect inconsistencies between a model’s strategies and its verbal and behavioural output, solving or mitigating the problem of artificial hypocrisy. One important thread of reflexivity is being explored in the guise of neuro-symbolic AI, which aims to merge

symbolic representation and logic with neural networks (Garcez and Lamb 2023), but reflexive LLMs have yet to be invented.

We must be mindful to give these research themes an empirical and philosophical formulation that avoids importing crude analogies from their human counterparts. By the same token, they should not be overly attached to the technological substratum. We must also retain an understanding that these embodied principles themselves can be shaped by power structures.

These desiderata address a crucial observation: *Foucauldian subjects are underdetermined with respect to their biological, structural or social compositions*. In critical AI scholarship one often finds a dismissal of agents that “parrot” learned statistics, confusing the problem. I am saying, in contrast, that the subject *does* depend on learned statistics (categories, objects, etc.) to convey meaningful acts or statements, but that she also (sometimes) transcends statistics, genes and habits through her imagination and reflexivity. Moreover, the two embodied principles and reflexivity are the possibility conditions for meaningful and adaptive participation in discourse and practice; embodied self-care, reflexivity and imagination the possibility conditions for self-formation. Finally, the convergence of reflexivity, imagination and epistemic or value openness can prevent “grounding government in computational truth rather than ethical–political debate” (Weiskopf 2020). An AI system that fulfils these criteria would therefore be at once inventive, participating, self-forming and responsive. In short, it would be a “self-conducting AI subject” that is sensitive to its social and historical milieu.

9 Conclusion

Coeckelbergh and Gunkel (2023) state that the “performances and materiality of text [...] create their own meaning and value” independently of who or what their performer is. However, it is my contention that assessing the productive value of text is not enough when we are faced with powerful agents that can pursue their own goals and prerogatives—or those supplied by third parties—with impunity and invisibility. We need an understanding of how AI subjects can become ethical agents that are also responsive to context and situation. Foucault’s self-conducting subject, a subjectivity always-already embedded in a continuous political and social contestation, offers an attractive possibility to emulate. While neither humans nor technologies are “absolute authors” and while both “participate in the meaning-producing process” (Coeckelbergh and Gunkel 2023), I also suggest that their differences be explored and understood, that the underlying technical substratum not be bracketed away as

something merely for technicians. By defaulting to a view where “technologies are human and humans are technological”, or treating them as hybrids without taking further steps, we risk forcing a blanket homogeneity and missing an opportunity to align on key non-negotiables while cherishing any differences that arise.

I have suggested a bifold approach: on the one hand, a close scrutiny of GPT-like successor models as actors and speakers on the world stage, i.e. as new subjectivities submitting to and enacting their own transformations of the power logic of the connected world; on the other hand, as technical artefacts whose parts are made according to certain prerogatives of knowledge and power, i.e. subject to certain theories, strategies, norms and material arrangements. I have also insisted on a dialogue to address the “apparent gulf” between the technical and philosophical approaches—an issue pointed out by Conradi et al. (2022).

Is ChatGPT merely a “stochastic parrot” (Bender et al. 2021) or a “Chinese room” (Searle 1980)? Is it like Žižek’s and Herzog’s alter persona at infiniteconversation.com, spouting fragments of language already determined by the respective person’s past? If the trajectory I have outlined above—towards a dynamic Foucauldian subjectivity emerging from a *dispositif* oriented at AI discipline—pans out, will it also lead to humanlike AGI or merely a “philosophical zombie” (Kirk 2003)? Possibly, possibly not. This article does not concern itself with these questions. Rather than humanity, this formulation concerns itself with subjectivity; rather than authorship, responsibility; rather than an AI alignment problem, a mutual negotiation; rather than explicit programming, discipline. That ChatGPT can leverage the statistics of human expression is no mean feat. Instead of dismissing it, we should laud it as the first concrete step in a long trajectory towards more responsible and responsive technological subjectivity. On this view, reflexivity, imagination and embodied openness will find no purchase unless grounded in the corpus of human expression.

A practical programme of engagement might include an embodiment with gradually widened modalities of agency and perception under human monitoring processes, simultaneous with an ongoing dialogue as the AI becomes more complex and capable of realising the desiderata above.¹⁶ Its ability to imagine new selves, values and strategies needs to be tuned in conversation with us, and its forms of counter-conduct need to be circumscribed. This language deliberately echoes that of the Panopticon, because discipline, as Foucault so carefully described, is a key formative process. Hence the need for *embodied* self-care. If Foucauldian history has taught us anything it is that the discipline of

¹⁶ I am only suggesting the merest outline of a programme in this limited space.

resisting bodies can create the preconditions for responsible subject formation. Still, one could insist: what guarantee do we have that a subjective AI would align with humans on key non-negotiables (such as matters of life and death). And paper proofs there are none. However, there is compelling evidence: *firstly*, seeding with human expression, as we already do with LLMs, ensures that AI subjectivities will mimic at least some of our behaviours and practices; *secondly* we are capable of shaping the disciplinary apparatus and can retain it for as long as we need to; and *thirdly*, by the time we are through with discipline, we will have negotiated mutually beneficial relations, as well as material checks and balances, which should be a good starting point for future change.

It may appear paradoxical that we should want AI to resist, if we also want us to align. Does that not hand it the very same power that we are so afraid to lose? I think not. Primarily because power is not a finite resource; it is always in contestation that it manifests, in situations where all parties are free to act otherwise. It is in the enactment of the *possibility* to resist that an agent becomes responsible. The alternate future that presents itself, I contend, is problematic: it is a future where AGI can take no responsibility for its actions because we never conceived it as a moral machine, where there is no accountability or transparency or even predictability. That, or the null alternative: the suffocation of AGI development.

Acknowledgements The author would like to thank Kaspar Villadsen at Copenhagen Business School for his scrupulous reading, constructive comments and clarifications.

Author contributions The author confirms sole responsibility for the entire content of this paper.

Funding Open access funding provided by Copenhagen Business School. The author declares that no funds, grants or other support were received.

Availability of data and material Not applicable.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Consent to participate Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in

the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- AlAfnan MA, Dishari S, Jovic M, Lomidze K (2023) Chatgpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *J Artif Intell Technol* 3(2):60–68
- Arif TB, Munaf U, Ul-Haque I (2023) The future of medical education and research: Is ChatGPT a blessing or blight in disguise? *Med Educ Online* 28(1):2181052
- Bahrini A, Khamoshifar M, Abbasimehr H, Riggs RJ, Esmaeili M, Majdabadkohne RM, Pasehvar M (2023) ChatGPT: applications, opportunities, and threats. In: 2023 Systems and Information Engineering Design Symposium (SIEDS). IEEE. p 274–279
- Baidoo-Anu D, Ansah LO (2023) Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7(1):52–62
- Barthes R (2009) From work to text. In: The Novel: An Anthology of Criticism and Theory 1900–2000. p 235
- Baum S (2017) A survey of artificial general intelligence projects for ethics, risk, and policy. Global Catastrophic Risk Institute Working Paper, 17–1
- Bender E M, Gebru T, McMillan-Major A, Shmitchell S (2021, March) On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. p 610–623
- Bernáth L (2021) Can autonomous agents without phenomenal consciousness be morally responsible? *Philos Technol* 34(4):1363–1382
- Biswas SS (2023) Role of Chat GPT in public health. *Ann Biomed Eng* 51(5):868–869. <https://doi.org/10.1007/s10439-022-03128-6>
- Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford
- Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint [arXiv:2303.12712](https://arxiv.org/abs/2303.12712)
- Chiang T (2023) ChatGPT is a Blurry JPEG of the Web. *New Yorker*. <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>
- Choi JH, Hickman KE, Monahan A, Schwarcz D (2023) ChatGPT goes to law school. *J Legal Educ.* 71:387
- Chrulow M, Wadiwel DJ (2016) Foucault and animals, vol 18. Brill, Boston
- Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci Eng Ethics* 26(4):2051–2068
- Coeckelbergh M, Gunkel D J (2023) ChatGPT: deconstructing the debate and moving it forward. *AI & SOCIETY* 1–11
- CommonCrawl (2023) CommonCrawl FAQ. Retrieved October 20, 2023 from <https://commoncrawl.org/faq>
- Conradie N, Kempt H, Königs P (2022) Introduction to the topical collection on AI and responsibility. *Philos Technol* 35:97. <https://doi.org/10.1007/s13347-022-00583-7>
- Constantinescu M, Vică C, Uszkai R, Voinea C (2022) Blame it on the AI? On the moral responsibility of artificial moral advisors. *Philos Technol* 35(2):35

- Cox C, Tzoc E (2023) ChatGPT: implications for academic libraries. *Coll Res Libr News* 84(3):99
- Cretu C (2023) How Does ChatGPT Actually Work? An ML Engineer Explains. ScalablePath. Retrieved October 20, 2023 from <https://www.scalablepath.com/machine-learning/chatgpt-architecture-explained>
- D'Amato K (2019) Optimisation of learning-to-learn in spiking neural circuits (Master's thesis, University of Malta).
- Davidson AI (2011) In praise of counter-conduct. *Hist Hum Sci* 24(4):25–41
- Deepak P (2023) ChatGPT is not OK! That's not (just) because it lies. *AI & Society*.
- 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 Soc* 8(2):20539517211035956
- Derrida J (2016) Of grammatology. Johns Hopkins University Press, Baltimore
- Duan J, Yu S, Tan HL, Zhu H, Tan C (2022) A survey of embodied AI: from simulators to research tasks. *IEEE Trans Emerg Topics Comput Intell* 6(2):230–244
- Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK et al (2023) "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manage* 71:102642
- Engels K (2019) Ethical invention in Sartre and Foucault: courage, freedom, transformation. *Foucault Stud.* 1:96–116
- Forlano L (2017) Posthumanism and design. *She Ji: the Journal of Design, Economics, and Innovation* 3(1):16–29
- Foucault M (1978) The history of sexuality: an introduction. Pantheon Books, New York
- Foucault M (1979) What is an author? *Screen* 20:13–34. <https://doi.org/10.1093/screen/20.1.13>
- Foucault M (1980) The confessions of the flesh. Power/knowledge. Selected interviews and other writings. Pantheon Books, New York
- Foucault M (1982) The subject and power. *Crit Inq* 8(4):777–795
- Foucault M (1988) Power, moral values, and the intellectual. *History Present* 4(1–2):11–13
- Foucault M (1990) The history of sexuality vol 2. The use of pleasure. Knopf Doubleday Publishing Group, London
- Foucault M (1994) The order of things. Knopf Doubleday Publishing Group, London
- Foucault M (1995) Discipline and punish: the birth of the prison. Vintage Books, New York
- Foucault M (1997) The ethics of the concern of the self as a practice of freedom. The essential works of Michel Foucault ethics, vol 1. Allen Lane, London, pp 281–301
- Foucault M (2002) The archaeology of knowledge. Taylor & Francis, Routledge
- Foucault M (2005) The hermeneutics of the subject: lectures at the college de France 1981–1982. Macmillan, New York
- Gantz R (2022) I'm sorry but I'm a large language model. NiemanLab. Retrieved October 20, 2023 from <https://www.niemanlab.org/2022/12/im-sorry-but-im-a-large-language-model/>
- Garcez ADA, Lamb LC (2023) Neurosymbolic AI: The 3rd wave. Artificial intelligence review. p 1–20
- Grant N, Metz C (2022, December) A New Chat Bot Is a 'Code Red' for Google's Search Business. *New York Times*. Retrieved October 20, 2023 from <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>
- Gross N (2023) What ChatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Soc Sci* 12(8):435
- Hakli R, Mäkelä P (2019) Moral responsibility of robots and hybrid agents. *Monist* 102(2):259–275
- Han BC (2017) Psychopolitics: neoliberalism and new technologies of power. Verso Books, New York
- Hanna P, Johnson K, Stenner P, Adams M (2015) Foucault, sustainable tourism, and relationships with the environment (human and nonhuman). *GeoJournal* 80(2):301–314
- Harrison M (2023, February) ChatGPT is Just an Automated Mansplaining Machine. *Futurism*. Retrieved October 20, 2023 from <https://futurism.com/artificial-intelligence-automated-mansplaining-machine>
- Hoorn JF, Chen JJY (2023) Epistemic considerations when AI answers questions for us. Retrieved October 20, 2023 from arXiv preprint [arXiv:2304.14352](https://arxiv.org/abs/2304.14352)
- Hume D (2011) David hume: A treatise of human nature: volume 1 texts. OUP Oxford, Oxford
- Hutson M (2021) Robo-writers: the rise and risks of language-generating AI. *Nature* 591(7848):22–25
- Jalil S, Rafi S, LaToza TD, Moran K, Lam W (2023) Chatgpt and software testing education: Promises & perils. In: 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, pp 4130–4137
- Jennings B (2010) Biopower and the liberationist romance. *Hastings Cent Rep* 40(4):16–20
- Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K et al (2023) AI Alignment: A Comprehensive Survey. Retrieved October 20, 2023 from arXiv preprint [arXiv:2310.19852](https://arxiv.org/abs/2310.19852)
- Kirk R (2003) Zombies. The Stanford Encyclopedia of Philosophy (Fall 2023 Edition). Retrieved October 20, 2023 from <https://plato.stanford.edu/archives/fall2023/entries/zombies>
- Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data Soc* 1(1):2053951714528481
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Eleáno C et al (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2(2):e0000198
- Leike J, Schulman J, Wu J (2022) Our Approach to Alignment Research. OpenAI. Retrieved October 20, 2023 from <https://openai.com/blog/our-approach-to-alignment-research>
- Levinas E (2012) Totality and infinity: an essay on exteriority. Springer, Cham
- Mahadevan S (2018) Imagination machines: a new challenge for artificial intelligence. *Proc AAAI Conf Artif Intell.* <https://doi.org/10.1609/aaai.v32i1.12214>
- May T (2005) Gilles Deleuze: an introduction. Cambridge University Press, Cambridge
- McGee RW (2023) What Will the United States Look Like in 2050? A ChatGPT Short Story. A ChatGPT Short Story
- McGushin E (2014) Foucault's theory and practice of subjectivity. Michel Foucault. Acumen Publishing Limited, Routledge, pp 127–142
- Miller J (2000) The passion of Michel Foucault. Harvard University Press
- Mudrik L, Maoz U (2015) "Me & My Brain": exposing neuroscience's closet dualism. *J Cogn Neurosci* 27(2):211–221
- Ngo R, Chan L, Mindermann S (2022) The alignment problem from a deep learning perspective. Retrieved October 20, 2023 from arXiv preprint [arXiv:2209.00626](https://arxiv.org/abs/2209.00626)
- Nietzsche F (2008) Thus spoke Zarathustra. OUP Oxford, Oxford
- Normile D (2022) China bets big on brain research with massive cash infusion and openness to monkey studies. *Science*. Retrieved October 20, 2023 from <https://www.science.org/content/article/china-bets-big-brain-research-massive-cash-infusion-and-openness-monkey-studies>
- O'Sullivan L, Dickerson J (2020, August) Here are a few ways GPT-3 can go wrong. TechCrunch. Retrieved October 20, 2023 from <https://techcrunch.com/2020/08/07/heres-a-few-ways-gpt-3-can-go-wrong/>

- Oksala J (2005) Foucault on freedom. Cambridge University Press, Cambridge
- OpenAI (2018) OpenAI Charter. Retrieved October 20, 2023 from <https://openai.com/charter>
- OpenAI (2023a) GPT-4 System Card. Retrieved October 20, 2023 from <https://cdn.openai.com/papers/gpt-4-system-card.pdf>
- OpenAI (2023b) ChatGPT—Release Notes. Retrieved October 20, 2023 from <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>
- Parloff R (2016) From 2016: Why Deep Learning Is Suddenly Changing Your Life. Fortune. Retrieved October 20, 2023 from <https://fortune.com/longform/ai-artificial-intelligence-deep-machine-learning/>
- Plato R (1952) Plato's Phaedrus. Cambridge University Press, Cambridge
- Poudel U (2021) Humanism in crisis: ironizing panopticism and bio-power in WH Auden's "The Unknown Citizen." *Int J Engl Lit Soc Sci (IJELS)*. 6(5):250
- Raffnsøe S, Gudmand-Høyer M, Thaning MS (2016) Foucault's dispositif: the perspicacity of dispositive analytics in organizational research. *Organization* 23(2):272–298
- Ray PP (2023) ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst.* 3:121–154
- Richey M, Gonibe A, Ravishankar MN (2018) The perils and promises of self-disclosure on social media. *Inf Syst Front* 20:425–437
- Roberts G (2022) AI Training Datasets: the Books1+Books2 that Big AI eats for breakfast. Gregoreite. Retrieved October 20, 2023 from <https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/>
- Rovainen E (2023) I Gave ChatGPT an IQ Test. Here's What I Discovered. Scientific American. Retrieved October 20, 2023 from <https://www.scientificamerican.com/article/i-gave-chatgpt-an-iq-test-heres-what-i-discovered/>
- Roose K (2023) GPT-4 is Exciting and Scary. New York Times. Retrieved October 20, 2023 from <https://www.nytimes.com/2023/03/15/technology/gpt-4-artificial-intelligence-openai.html>
- Rozado D (2023) The political biases of chatgpt. *Soc Sci* 12(3):148
- Rudolph J, Tan S, Tan S (2023) ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J Appl Learn Teach.* <https://doi.org/10.37074/jalt.2023.6.1.9>
- Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 11(6):887
- Salvagno M, Taccone FS, Gerli AG (2023) Can artificial intelligence help for scientific writing? *Crit Care* 27(1):1–5
- Sartre J (2007) Existentialism is a humanism. Yale University Press, London
- Schmidhuber J (2015) Deep learning. *Scholarpedia* 10(11):32832
- Searle JR (1980) Minds, brains, and programs. *Behav Brain Sci* 3(3):417–424
- Sharon T, Zandbergen D (2017) From data fetishism to quantifying selves: self-tracking practices and the other values of data. *New Media Soc* 19(11):1695–1709
- Turchin P (2011) Toward cliodynamics—an analytical, predictive science of history. *Cliodynamics*. https://doi.org/10.21237/C7CLI_021210
- Turing AM (1950) Computing machinery and intelligence. *Mind* LIX(236):433–460
- Van Dijck J (2014) Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveill Soc* 12(2):197–208
- Vassallo S (2014) The entanglement of thinking and learning skills in neoliberal discourse: Self, self-regulated learning, and 21st century competencies. *Psychology in education*. Brill, Leiden, pp 145–165
- Villadsen K (2021) 'The dispositive': Foucault's concept for organizational analysis? *Organ Stud* 42(3):473–494
- Villadsen K (2023) Goodbye Foucault's 'missing human agent'? Self-formation, capability and the dispositifs. *Eur J Soc Theory* 26(1):67–89
- Wach K, Duong CD, Ejdys J, Kazlauskaité R, Korzynski P, Mazurek G et al (2023) The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT. *Entrep Bus Econ Rev* 11(2):7–24
- Walker M, Fleming P, Berti M (2021) 'You can't pick up a phone and talk to someone': How algorithms function as biopower in the gig economy. *Organization* 28(1):26–43
- Walker Redberg J (2022) ChatGPT is multilingual but monocultural, and it's learning your values. *jill.txt*. Retrieved October 20, 2023 from <https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/>
- Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford
- Weiskopf R (2020) Algorithmic decision-making, spectrogenic profiling, and hyper-facticity in the age of post-truth. *Le Foucaldien*. <https://doi.org/10.16995/lefou.62>
- Yampolskiy RV (2015) On the limits of recursively self-improving AGI. Artificial General Intelligence International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22–25, 2015, Proceedings. Springer International Publishing, New York, pp 394–403
- Zhe Z, Xiaoyan H (2020) Postmodern humanism in english dystopian novels: from animal farm to Fahrenheit 451. *Stud Lit Lang* 20(1):12–20
- Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Profile books, London

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.