

**“i am a stochastic parrot, and so r u”:  
Is AI-based framing of human behaviour and cognition a  
conceptual metaphor or conceptual engineering?**

**Abstract**

Understanding human behaviour, neuroscience and psychology using the concepts of ‘computer’, ‘software and hardware’ and ‘AI’ is becoming increasingly popular. In popular media and parlance, people speak of being ‘overloaded’ like a CPU, ‘computing an answer to a question’, of ‘being programmed’ to do something. Now, given the massive integration of AI technologies into our daily lives, AI-related concepts are being used to metaphorically compare AI systems with human behaviour and/or cognitive abilities like language acquisition. Rightfully, the epistemic success of these metaphorical comparisons should be debated. Against the backdrop of the conflicting positions of the ‘computational’ and ‘meat’ chauvinisms, we ask: can the conceptual constellation of the computational and AI be applied to the human domain and what does it mean to do so? What is one doing when the conceptual constellations of AI in particular are used in this fashion? Rooted in a Wittgensteinian view of concepts and language-use, we consider two possible answers and pit them against each other: either these examples are *conceptual metaphors*, or they are attempts at *conceptual engineering*. We argue that they are conceptual metaphors, but that (1) this position is unaware of its own epistemological contingency, and (2) it risks committing the “map-territory fallacy”. Down at the conceptual foundations of computation, (3) it most importantly is a misleading ‘double metaphor’ because of the metaphorical connection between human psychology and computation. In response to the shortcomings of this projected conceptual organisation of AI onto the human domain, we argue that there is a semantic catch. The perspective of the conceptual metaphors shows avenues for forms of conceptual engineering. If this methodology’s criteria are met, the fallacies and epistemic shortcomings related to the conceptual metaphor view can be bypassed. At its best, the cross-pollution of the human and AI conceptual domains is one that prompts us to reflect anew on how the boundaries of our current concepts serve us and how they could be approved.

**Authors**

Warmhold Jan Thomas Mollema. <https://orcid.org/0009-0001-3269-6837>.

Thomas Wachter. <https://orcid.org/0000-0003-0561-969X>.

**Conflict of interest**

The authors have no conflict of interest to declare and received no funding for writing this article.

## 1. Introduction:

### **The conceptual muddle of the human and the computational**

In a recent *New Yorker* article, Angie Wang (2023) sketches a story of a mother that perceives parallels between her toddler's process of learning language and the 'stochastic parroting' of LLMs (Bender et al., 2021). When she witnesses her son stringing together vowels and calling many different objects 'cat', she directly employs AI-language to make sense of his behaviour. She asks: "Aren't we after all just a wetware neural network, a complex electrochemical machine?". When she corrects him calling everything 'cat', she asks "So is this supervised learning?". In short: "Is my Toddler a Stochastic Parrot?". This functionalist explanation is gaining credit today. Like OpenAI's CEO Sam Altman (2022) once tweeted: "i am a stochastic parrot and so r u". Wang's mother-protagonist, at a moment of despair, comes to embrace this description: "ChatGPT is a stochastic parrot and so are we." So in the case of the 'stochastic toddler', human development and language use, as a whole, is understood in terms of the computer science and mathematics of machine learning models.

This is an example of a bigger tendency. Given the increasing integration of AI-based systems in human life, AI has become a more present image to use in order to describe human behaviour and cognition. The tendency, which can be understood as a new iteration of the computer-metaphor in cognitive science and psychology (Baria & Cross, 2021; Brette, 2022; Piccinini, 2009; Shagrir, 2006), is receiving philosophical attention. In particular, the debate revolves around whether AI should be allowed access to the concepts in the human arsenal, from both a theoretical and an ethical perspective. For instance, Floridi and Nobre (2024) discuss how the use of AI-related concepts in cognitive science can obscure important aspects of our cognition. In turn, Mitchell (2024a) notes that the use of the metaphors used to describe AI systems affect the way we interact with them, how much we trust them and also the different ways we regulate them. In this context, we raise the driving question of this work: is the conceptual constellation of the computational and AI applicable to the human domain and what does it mean to do so? How should these attempts made in science and society at understanding human behaviour, psychology and neuroscience in terms of AI be understood methodologically? We consider two possible answers and pit them against each other: either these examples are *conceptual metaphors*, or they are attempts at *conceptual engineering*.

In Wang's literary example, the temptation of understanding a toddler in terms of AI is well sketched out, but is immediately quenched by asserting the irreplaceability of human bodily bonds for development. In the article's finale, the mother asserts that "what we say isn't weighted by probability". In other words, by recognizing the irreplaceable aspects of the human, she believes these concepts do not capture what humans are or do.

On the other hand, defending human specialness is regarded as a form of 'meat chauvinism' (Aaronson, 2024; Machine Learning Street Talk, 2024b, 2024a). In other words, the parallels between human and AI behaviour makes room for explaining the former in concepts that belong to the latter. Stories like Wang's appeal to the contrary, but make no convincing argument for why humans would be *fundamentally* any different than machine learning systems. As quantum physicist

Aaronson ridicules this position: “You wanna stomp your feet and be a meat chauvinist?” (Aaronson, 2024). It seems like a face-off between ‘computational chauvinism’ (Piccinini, 2006) and ‘meat chauvinism’ is imminent.

To answer our main questions, the remainder is structured as follows. Section 2 historicizes the application of the metaphors of ‘machine’ and ‘computer’ to the human domain and proposes two possibilities for viewing the deployment of AI concepts: the conceptual metaphor view and conceptual engineering view. Section 3 discusses conceptual metaphor theory and conceptual engineering to deepen our understanding of both. Section 4 covers the conceptual metaphor view’s good fit to the stochastic parrot case. This is where our main argument comes in. We argue that, at a functional level of explanation, (1) this position is unaware of its own epistemological contingency, and (2) it risks committing the “map-territory fallacy”. Finally, at the level of the conceptual foundations of computation, (3) we show this computational style of explanation is a misleading ‘double metaphor’ because of the metaphorical connection between human psychology at the root of computation. In response to these shortcomings, we argue in section 5 that conceptual metaphors’ perspectives can still provide cues or avenues for forms of conceptual engineering. If this methodology’s criteria are met, the conceptual metaphor view’s epistemic shortcomings can be bypassed.

## **2. The machine conception of organism and AI: In between metaphors and conceptual engineering**

Few notions in biology have been so pervasive as the machine conception of the organism (MCO). Formulated by Descartes in the seventeenth century, the MCO is based on the metaphorical re-description of organisms as machines in order to understand them (Nicholson, 2013). For instance, in 1665 the Danish anatomist Nicolaus Steno boldly argued that if we want to understand how and what the brain does, we should take it apart and, with that, view it as a machine (Cobb, 2020). From there, the history of theorising organisms became entangled with technological development, constantly comparing ourselves, metaphorically, with the most advanced technology of the time (Birhane, 2021).<sup>1</sup>

Since the advent of the computer in the 1950s, the study of brain and cognition has been dominated by approaches and concepts from information theory and computer science (Floridi & Nobre, 2024). Concepts like “information processing”, “coding”, “computing”, and “algorithm”

---

<sup>1</sup> In the seventeenth century, it referred to clockworks—intricately calibrated parts working together as a single, integrated unit. By the eighteenth century, it described steam engines, which consumed energy through combustion to perform work and generate heat. In the nineteenth century, the term extended to chemical factories, where machines coordinated and managed complex networks of interconnected chemical reactions (Nicholson, 2013).

frame the brain and mind as computational information processing systems (Brette, 2022; Colombo & Piccinini, 2024; Floridi & Nobre, 2024). This new version of the MCO, has been instrumental to the development of cognitive science, psychology and AI (Felin & Holweg, 2024).<sup>2</sup>

The MCO stands at the core of AI which regards cognition as a general form of computation. For example, according to the field’s founders (convening at the Dartmouth Conference in 1956), their goal was to “proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al., 2006). Moreover, McCulloch and Pitts’ first mathematical model of the neuron –foundational to modern AI–, was directly inspired by the analogy of the brain as a computer in which neurons were modelled like logic gates (Chirimuuta, 2021).<sup>3</sup> Despite criticism of the metaphorical connection between brains and computers/AI, this conceptual framework remains highly regarded in cognitive science. Some argue that it provides valuable conceptual and formal tools for theory development and enables meticulous assessment of computational feasibility (Van Rooij et al., 2024). As Margaret Boden (1988) puts it: “Computers as such are, in principle, less crucial for cognitive science than computational concepts are.”

Complementarily, we make the negative argument that today’s understanding of human behaviour and psychology in terms of AI are *conceptual metaphors*. Leaving this implicit in the prevalent discourse has significant downsides. Appeal to metaphor doesn’t explain its object of description, but rather emphasises certain convergences at the cost of suppressing divergences. Our corresponding positive argument consists in how the methodology of *conceptual engineering* could be applied to these domains, based on the cues given by new AI metaphors.

But first, we set up both the conceptual metaphor and conceptual engineering views, considering them as equal contenders.

Understanding MCO’s novel manifestations as *conceptual metaphors* is to view them as systematic projections of one conceptual domain onto another for highlighting similarities between both domains (Kövecses, 2017). Put differently, the conceptual domain of the computational is structurally organised and deployed to put an analogy to work in the domains of life, mind and humanity. Call this the conceptual metaphor view:

*CM*: Using computational concepts and AI to explain human behaviour, cognition and psychology should be understood as *the projection of the former conceptual domain onto the latter domains, which highlights similarities, while suppressing differences between them*. This is an epistemological comparison because it structures the way in which we can know the target domain.

---

<sup>2</sup>For example, in the influential book that initiated the field of Cognitive Psychology, Ulric Neisser states: “The task of a psychologist trying to understand human cognition is analogous to that of a man trying to discover how a *computer has been programmed*” (Neisser, 1967, p. 6, emphasis added).

<sup>3</sup> It is important to mention that they believed that, under certain conditions, the brain was literally a computer (Chirimuuta, 2021).

Alternatively, these attempts can be viewed as *conceptual engineering* – making changes to the intensions and extensions of concepts in order to change their content, usage or domain of reference (Cappelen, 2018). That is: understanding for example ill-defined concepts like ‘mind’ (Mitchell, 2024b) in terms of AI prompts the crafting, amelioration or elimination of concepts belonging to the conceptual constellation of the mind. Call this the conceptual engineering view:

*CE*: Using computational concepts and AI to explain human behaviour, cognition and psychology should be understood as shifting conceptual boundaries for *extending, eliminating or refining existing concepts to better fit reality*. Extant concepts are altered, replaced or eliminated because the conceptual resources of the computational and AI provide better models for these domains.

Returning to the ‘stochastic toddler’ and similar examples, we observe an initial tension between *CM* and *CE*. First, on *CM*, we can identify a core metaphor: HUMAN INTELLIGENCE is ARTIFICIAL INTELLIGENCE, or to paraphrase: INTELLIGENCE is ARTIFICIAL.<sup>4</sup> As core conceptual metaphor, this underlies instances like the ‘stochastic toddler’, which are reversals of the “‘LLM as individual mind’ metaphor” (Mitchell, 2024a). *Prima facie* this is a moderately/strongly epistemically successful conceptual metaphor (see section 3). Conversely, in *CE*, we can see the drawing of conceptual interconnections from HUMAN to AI as cues for the conceptual re-engineering of human domains via computational concepts.

In this paper’s remainder, we thoroughly set up both views (section 3) and pit both views against each other in an analysis of the explanatory human-AI nexus (sections 4 and 5). Our hypothesis is that cases like ‘stochastic toddler’ converge to *CM* rather than to *CE*, which epistemically limits the fruits that can be reaped from understanding human domains in terms of AI: it is not necessarily so that the computational and AI provide their best and only model. Conceptual practices, like thinking of the brain in terms of computers, are shaped by the way language is used and how language-users are initiated into conceptual practices (Wittgenstein, 2009, §§1-37; Mollema, 2024a). The practice of stressing similarities turns the framing of organismic concepts as computational into a seemingly *metaphysical* one, while, as we will show, it’s at best a contingent epistemological perspective. The logico-grammatical conceptual organisation of the computational is one ‘we’ – those bombarded with computers and AI – come to find attractive, but for which alternatives are possible. We are socioculturally initiated into the usefulness of talk, models, and explanations based on computers/AI to make sense of human domains. This practice has proliferated to the point that it has become intuitive for many language-users. Differently stated, this particular use of concepts is in the process of *fossilisation*, becoming a certainty for many of us (Wittgenstein, 1975, §474). But unbeknownst to them, any language-game can in principle be played differently, while still satisfying its reason for being played: in our case the scientific and lay understandings of the human domain. However, if we are aware of the differences these metaphors

---

<sup>4</sup> Following (Lakoff & Johnson, 2003), strings that indicate conceptual metaphors are written in upper-case, connected by lowercase verbs.

suppress and the oversimplifications they engender, current AI advancements offer a unique opportunity for conceptual engineering. *CE* is a better fit because it doesn't enforce the computational view, but rather enables the re-evaluation of established concepts in terms of phenomena that push the boundaries of their applicability. We can use this as input for social processes of ethically considering which concepts to go on with (Queloz, 2021; Queloz & Cueni, 2021).

### 3. Conceptual metaphor theory and conceptual engineering

#### 3.1. Conceptual metaphors

Max Black's (1955) 'interaction theory of metaphor' revived metaphor theory and conceptualised a metaphor as having a *principal* subject ('target domain') and a *subsidiary* subject ('source domain'). In the case of the metaphor 'that man is a wolf', the principal subject is 'that man' and the subsidiary subject is 'a wolf'. When the metaphor is evoked, it "selects" or "filters" aspects from a target domain by projecting the wolf-"system of commonplaces" onto the source domain (man) (Black, 1955, pp. 73–75).<sup>5</sup> Every metaphor is "the tip of a submerged model" (Black, 1977, p. 445): a picturing of something from a certain perspective, with explanatorily valuable expressive power. A successful metaphor discloses the target domain's "intelligible structure" (Miller, 1979). So a metaphor's provided model is an analogical organisation of two conceptual domains for stressing certain similarities, while suppressing certain differences.<sup>6</sup> What is a *conceptual* metaphor then? Conceptual metaphor theory, developed by George Lakoff and Mark Johnson (1981; 2003) builds upon Black's conceptualization. A conceptual metaphor is a *set of correspondences* between source domain and target domain (Kövecses, 2017; Lakoff, 1993; Lakoff & Johnson, 2003) that enables a structural understanding of "coherent organization[s] of human experience" (Kövecses, 2008). Common examples are 'TIME is MONEY', 'LOVE is a JOURNEY', 'ARGUMENT is WAR', 'UP is BETTER' or 'THEORY is a BUILDING', and there are many ways in which the resemblance between the source and target domain is spelt out in everyday parlance. E.g., many variants of expressions like 'I'm *short* on time', 'she *attacked* the premise', or 'the neuron is *fundamental* to neuroscience' are commonly used. Conceptual metaphors, therefore, make two domains 'emphasise and resonate' (Black, 1977, pp. 439-440). Furthermore, conceptual metaphors aren't arbitrary: they "are grounded in systematic correlations within our experience" (Lakoff & Johnson, 2003, p. 61). They go over and above individual metaphors, because they express a descriptive invariance that can be constructed between two domains. It is a regular pattern of similarities running through our everyday parlance and thought.

---

<sup>5</sup> Of course, Black's view was not without critics. It is out of scope however to engage in the debate on this matter.

<sup>6</sup> This invitation to take a certain perspective is more than merely aesthetic: a "distinctive mode of achieving insight", enabling us "to *see new connections*" (Black, 1962, pp. 223; 236-237).

A metaphor's perspective can be more or less epistemically successful. Metaphors also play scientific roles (Taylor & Dewsbury, 2018; Rivadulla, 2006); think of the atom, electricity as a current, the heart as a pump, etc.. To explain how metaphorical projections can be inescapably flawed, while still being epistemically successful, Nikola Kompa (2021, p. 41) has proposed criteria to measure the epistemic success of metaphors with:

- whether the entities in the metaphor exist;
- if inferential patterns are present between domains;
- the extent of the disanalogies between the domains;
- the theoretical merits of the metaphor's perspective;
- the extent to which necessary properties of the target domain are obscured;
- whether it engenders new metaphors.

Intuitively, whether or not these criteria are met, a conceptual metaphor at least discloses knowledge of some target domain in a structurally successful way: it encompasses a set of correspondences, while a 'normal' metaphor is tailored to a specific context.<sup>7</sup> However, conceptual metaphors also have limits. We may talk about love *as if* it's structured like a journey, but love also has dispositional, microbiological and psychological aspects, unshared with journeys. In short, "we prefer to exploit *contingent* properties, not *defining* properties, metaphorically" (Kompa, 2021, p. 34). Metaphor is "a seeing-as which occurs only while I am actually concerning myself with the picture as the object represented" (Wittgenstein, 2009, II, §199); a framing of one concept by another that depends on an agent implementing said perspective of the target concept.

With this exposition of conceptual metaphors in mind, we turn to a discussion of conceptual engineering.

### 3.2. Conceptual engineering

Conceptual engineering is a longstanding philosophical practice (Löhr, 2024): proposing new representational devices, disposing of old ones, or redrawing the boundaries of existing concepts. Changing our networks of concepts not only concerns tailoring concepts to the 'joints of nature', but also requires ethical reflection, i.e., the evaluation of concepts. Conceptual ethics is the normative reflection on whether the concepts we have are the concepts we *should* have, if the boundaries they lay down for us to go on with about our lives suffice for our needs (Queloz, 2021). And it is about how the concepts we use now are related to "us" as a social group, distinct from other social groups (Mollema, 2024a; Queloz & Cueni, 2021). While it is helpful to distinguish engineering and ethics, the conceptual engineer often fulfils both roles (Löhr, 2024), and for matters of simplicity, we treat the methodology of conceptual engineering as relating to both logical-grammar and ethics.

The contemporary debate on conceptual engineering isn't preoccupied with achieving so-called 'conceptual hygiene', but rather with how conceptual engineering ties in with forms of social

---

<sup>7</sup> Calling a man a wolf clearly works differently than thinking time analogously to money. For example, there are numerous expressions that are part of TIME is MONEY, such as 'I'm short on time' or 'time is scarce', etc.

change (Hopster et al., 2023). For example, Hopster & Löhr (2023, p. 8) argue that conceptual disruptions are the roots of conceptual engineering. An unsettlement of how a concept is collectively understood is what re-engineering our conceptual network addresses. In the context of AI that concerns us, conceptual engineering responds to the wind of societal disturbances that AI brought along.

Simplified, these disturbances come in three forms:

- ⇒ *gaps*: the conceptual tools are lacking for subsuming new phenomena (Veluwenkamp et al., 2024). For example, a natural anomaly is observed, like a new type of particle, or unforeseen circumstances occur, like the invention of a new technology.
- ⇒ *mismatches*: a conceptual delineation is deemed unfit for its current purpose(s). For example, using the concepts of ‘soul’ and ‘mind’ to account for human cognition and psychology was once suitable, but ended up in mismatch with mainstream scientific perspectives. What Hopster & Löhr (2023, p. 10) call ‘misalignment’ is a specific case of a mismatch, where “a given concept that is entrenched in a joint conceptual scheme is insufficiently aligned with the overall goals (prudential, moral, etc.) of the agents who deploy it”.
- ⇒ *improvements*: a new conceptualization is proposed to supersede the status quo. Such a proposition can serve to address a ‘conceptual overlap’ (Hopster & Löhr, 2023). Think of the semantic recharge of the concept of ‘gender’ in the 1960’s and its subsequent differentiations that allowed for novel granular expressions of identity.

Conceptual gaps, mismatches or improvements arise via technological, socio-cultural or psychobiological change. These changes can be gradual and general, as well as *ad hoc*, because historically philosophers, scientists and politicians have made a wealth of different changes to our conceptual network. To complete the introduction of conceptual engineering, we map gaps, mismatches or improvements to how (1) we can change existing concepts for the better; (2) introduce new concepts; or (3) eliminate existing concepts.

(1) Changing existing concepts for the better is termed ‘conceptual amelioration’ by Haslanger (2020) and ‘adaptation’ by Hopster & Löhr (2023). As Haslanger (2020, p. 230) says: “we should seek not only to elucidate the concepts we have, but aim to improve them in light of our legitimate purposes”. These improvements come in *epistemic* as well as *semantic* forms that respectively concern changing concepts’ knowledge structures and acquisition, and changing a concept’s partitioning of logical space. Epistemic amelioration targets how and what we can know about our representational tools, whereas semantic amelioration adjusts conceptual intensions and extensions. We can epistemically (i) refine a concept to gain better knowledge of what it represents to us and (ii) improve our experiential access to the concept’s informational content (Haslanger, 2020, p. 242). Semantically, we can (iii) socially reconstruct concepts so that concept-users can deploy them; (iv) pragmatically rearrange the terms of coordination of a concept (when, where and how it is appropriate to use it); and (v) change the ethical significance of a concept’s logical space.



In short, amelioration/adaptation can alternatively be described as forms of rational preservation of existing concepts.

Consider the concept ‘spider’ as an illustration. We might want to narrowly define ‘spider’ to better understand what it represents to us and avoid the confusion with, e.g., Japanese spider crabs. The scientific denomination *Araneae* does this by epistemically refining spider (*cf.* (i)). However, users of the word spider may have never encountered one. In that case we could improve their representational access to what using ‘spider’ is for by showing them one (*cf.* (ii)).<sup>8</sup> We could also socially rearrange spider’s conceptual delineation to exclude the property of having eight legs. This would be apt were a new species of nine-legged spiders to be discovered (*cf.* (iii)). Now imagine a more scientifically stringent society; in that case, the terms of coordination of ‘spider’ could be changed: we can determine that it denotes the ‘order’ abstraction and is no longer fit to be used to describe ‘species’ abstractions like *Tegenaria domestica* or *Poecilotheria metallica* (*cf.* (iv)). Finally, ‘spider’ might change in ethical significance and come to be seen or used as a morally harmful one, for example because of collective associations with a terrorist group called ‘The Spiders’.

(2) Fabricating new concepts – called ‘lexical expansion’ (Cappelen, 2020) or ‘introduction’ (Hopster & Löhr, 2023) – is more straightforward, but hard to do sustainably. It resembles neologisms, which can either be new views on existing concepts, or present fully novel concepts. For example, recent conceptual introductions regarding social media are ‘enshittification’ and ‘doomscrolling’. Both demarcate behaviours and processes that were either previously non-existent or not concisely articulable. A new item expands our collective lexicon that is genuinely distinctive (Cappelen, 2020) – not just some different word for the same concept, like ‘stool and chair’, which are synonyms, or the trio of ‘Venus, morning star and evening star’, which are denotations of the same object, but which emphasise different contextual properties. Cappelen(2018, pp. 123–128) writes: “An expression can have cognitive and emotive effects over and beyond (and in some sense independently of) any of its semantic and pragmatic properties”. In short, the introduced concept is informationally, cognitively, emotively and referentially novel.<sup>9</sup> Of course, the delineation of concepts is not necessarily *exact*, and conceptual overlap need not be a problem, nor is one of these angles fundamental for the others (Wittgenstein, 2009, §§76-77).

(3) Lastly, there is *elimination*. Famously, the Churchlands propagated the program of ‘eliminativism’. It expected advancements in neuroscience would make folk psychological concepts obsolete, and fit for elimination and replacement by scientifically underpinned concepts (Churchland Smith, 1980; Bennett & Hacker, 2003). Steffen Koch’s (2023) take on elimination – ‘explanatory eliminativism’ – is more recent. It holds that revisions can be taken too far, and that existing concepts can prove obsolete by lacking explanatory value. Koch (2023, p. 2141) argues there is a

---

<sup>8</sup> Note however that this example depends on the fact that spider is a strongly referential concept – i.e., tied to beings in the world. We can give ostensive definitions to ground purely referential concepts. For other types of concepts, other forms of improving experiential access are possible.

<sup>9</sup> As the previous examples tried to show, we have other ways of linking and distinguishing related words, e.g. via synonymy and taxonomy.

“strong pull” towards elimination because topical limits for revision are overrated. To illustrate, consider the historical examples of Bergson’s *‘élan vital’* and Stahl’s ‘phlogiston’. The terms were coined to explain something about life and matter respectively. However, once superseded by scientific advancements, the concepts’ expressive powers yielded no additions in explanatory value to theories making use of them. *Ergo*, they have been eliminated from the collective conceptual framework (despite becoming objects of study for the history of philosophy and science).

Now we promptly relate elimination, introduction and amelioration to the three forms of conceptual disruption. *Gaps* can be addressed by the introduction of new concepts or the extension of existing concepts’ scope, i.e., gaps can be filled or covered. *Mismatches* can be resolved by providing new delineations of existing concepts – or deliberately allowing for a new blurry overlap – and secondarily by introducing new concepts to cover the area of controversy. Finally, *improvements* to our conceptual network can be made by introducing new concepts to fortify our arsenal, and also by finetuning existing concepts, which is a form of adaptation. But what about elimination? Elimination is a curious case, as it can resolve mismatches *and* lead to improvements by removing dysfunctional concepts. However, it can only address a gap if the eliminated concept is replaced with something else, or if the elimination shows that an apparent gap was unjustified.

### 3.3. A Wittgensteinian conceptual intermezzo

A Wittgensteinian view of concepts prefigures how we evaluate conceptual interaction (*CM*) and conceptual emergence, change, or disappearance (*CE*).<sup>10</sup> Words are part of the human life form’s language-games (*Sprachspielen*). Words signify concepts while playing strongly context-dependent roles to achieve certain ends (Wittgenstein, 2009, §§7-18). In language-games, concepts are communicatively arranged sufficiently for some context-dependent purpose – and this is the cause of why the game is played. Concepts serve the language-game’s *point*, its pragmatic application (Wittgenstein, 1975, §474). Word-use points to meaning (*this* particular contextual arrangement of concepts). Therefore reasons, inequivalent with the cause of playing the language-game like *this*, can be given for why concepts are part of our life form’s contextual tapestry. For example, uttering ‘apple’ can denote different objects that instantiate different concepts (fruit, company, etc.), but also fits metaphorical employment (‘comparing apples and oranges’). Furthermore, the *need* for denotative concepts like ‘apple’ is undergirded by language-games such as pointing, describing, referring, and so on. We have come to have *these* concepts because of underlying needs to use them thus and so.

Defining concepts figuring in everyday life is often hard, impossible, or misguidedly futile. As Wittgenstein (2009, §569) says: “concepts are instruments” that determine how we can go about and while some may fit the same purpose, others take “more time and trouble than we can afford.”

---

<sup>10</sup> This is a deliberately shallow presentation of Wittgenstein’s ideas. We show what we need in the remainder, because that suffices for our purposes. For a more in-depth introduction Wittgenstein’s work in the context of neuroscience, see (Bennett & Hacker, 2003), in the context of social sciences, see (Pitkin, 1972), and for Wittgenstein’s epistemology in *On Certainty* see (Mollema, 2024a).

Technical domains are familiar with jargon with agreed upon definitions, but concepts that structure everyday life often completely evade definition. According to Wittgenstein (Pitkin, 1972, p. 65) the urge for definition leads to philosophical confusions, because “a complicated net of similarities that overlap and intersect” is what connects instances of concepts, “sometimes fundamental [...], sometimes [...] in details” (Wittgenstein, 2009, §66). *Family resemblance* rather than definition makes many concepts usable for us. Regarding defining a concept, i.e., drawing a boundary around it, Wittgenstein (2009, §69) reminds us that

We don't know the boundaries because none have been drawn. To repeat, we can draw a boundary – for a special purpose. Does it take this to make the concept usable? Not at all! Except perhaps for that special purpose.

For metaphors and conceptual engineering, this insight is of the essence. Defining concepts can only approximate concepts' informational and logico-grammatical contours, whereas the idea of jargon represents the drawing of “a boundary – for a special purpose”. Metaphors approximate one conceptual domain through the lens of another conceptual domain. Conceptual engineering modifies a concept's intension (its semantic reach) and extension (its domain of application) which depends, intuitively, on having some clarity with respect to a conceptual scheme's received definition.

This view is neither incompatible with metaphor theory, nor with practising conceptual engineering, but does pre-empt any ‘great expectations’ like metaphors showing us concepts’ ‘unchangeable essences’, or conceptual engineering supplying us with ‘perfectly clear representational devices’. If language needs fixing, we can only hope to do so *in situ* and *in medias res*, rather than *in vitro* and *ab ovo*.

#### 4. The limits of the conceptual metaphor view

Against this Wittgensteinian backdrop, we scrutinize the epistemological limits of applying *CM* to AI-based explanations of the human domain. Appeal to computational concepts to explain human behaviour and mind (e.g., in the stochastic toddler) enacts a conceptual projection. We argue this computational projection is fundamentally metaphorical, because it applies two non-identical conceptual domains to each other. Therefore, together with its structural ubiquity, *CM* applies to this schema of explanation.

In what follows, we unpack *CM*'s applicability to projecting computational concepts on human behaviour, cognition and psychology by showing how it captures similarities to front, while suppressing dissimilarities. Secondly, we substantiate the view's major shortcomings. (4.1.) We show this conceptual organisation is an epistemologically perspectival one, grounded in socio-political language-use. Furthermore, (4.2.) *CM* risks committing the ‘map/territory fallacy’: mistaking properties of the projective model for aspects of the target system. Finally, (4.3.) the explanation of INTELLIGENCE is ARTIFICIAL is a *double metaphor*: a metaphorical framing of the compu-

tational as cognitive. The technically charged instances of *CM* (layer 1) are prefigured by a modelling move at the foundation of computer science: the Turing machine is underpinned by a metaphorical connection between human calculation and mechanical computation (layer 2).

Conceptual metaphors build epistemic bridges through a set of correspondences between source and target domains, resulting in *emphasis* and *resonance*. The correspondences yield an analogical model for understanding and explaining the target system: a form of explanatory abstraction through analogy. It allows for understanding a complex, opaque system via similarities with a better understood system (Chirimuuta, 2020).<sup>11</sup> This process *abstracts* because for metaphors to resonate and, ultimately, be epistemically successful, highlighted similarities are maximally salient and disanalogies have been minimised. The epistemic tension here is that while metaphors help understand what's otherwise obscure, their imagery guides one's view and favourably spotlights a particular, incomplete perspective of the target system (Möck, 2022). Because of the success of AI-/computer-related metaphors, we can rightly ask what other vocabularies would supply better framings. Drawing parallels between AI systems and cognitive processes engenders new insights and fosters the development of hypotheses (Van Rooij et al., 2024; Kompa, 2021). But AI-related metaphors' abstraction has three clear limits.

#### 4.1. The conceptual metaphor view's epistemological contingency

The Wittgensteinian approach to concepts leads us to an epistemological perspectivism regarding the modelling capacities of conceptual networks. The seemingly logical connection between mechanical and organismic concepts is at best an *epistemological perspective* and a logico-grammatical conceptual organisation 'we' – those bombarded with computers and AI – come to find attractive. It is a *certain* "net" we cast out over the world (Wittgenstein, 1922, 6.342). But alternatives to this organisation are available.

As conceptual framing it guides the explanation of human domains, but is not 'set in stone'. As model, it cannot lay claim to 'the true contours of natural kinds'; instead, it is a more or less coarse grained approximation of the target domain in terms of a source domain. For when we talk of the brain as 'implementing algorithms', or humans 'employing language models', this concerns an approximation at a *functional level*, the wording of which applies a conceptual framing that cleaves our momentous understanding from more fine grained, alternative descriptions. The *pre-framing* of the image of the computational captivates us and keeps us captive *within* the linguistic range of the computational (Wittgenstein, 2009, §115). The spotlight on a functional level of abstraction, modelable by Turing-machines, places framings in the dark – just like talking about human behaviour with everyday concepts of folk psychology cleaves one from using neuroscientific terms for equiv-

---

<sup>11</sup> 'Better understood' is relative to a metaphor's target group, e.g. for scientists or lay-people. Given that most modern AI systems are considered to be unexplainable "black-boxes" (Rudin, 2019), this might sound like a contradiction. However, the comparison is usually made at a high level of description (Birhane & McGann, 2024), which allows the easy conceptual transfer between the source and the target system.

ocal description. Furthermore, what we learn from Queloz & Cueni (2021) is that conceptual practices of making sense of something in a certain way are politically maintained. The political maintenance of the concept-use of a certain ‘we’ (a social grouping sharing a foundational worldview) proceeds via the human control of initiation into language-games. To conclude, it is not that changing the rules of logical grammar is impossible. On the contrary, Wittgenstein (2009, §132) emphasised that it may “appear as if we saw it as our task to reform language. Such a reform for particular practical purposes, an improvement in our terminology designed to prevent misunderstandings in practice, may well be possible.” It is only that, paraphrasing Wittgenstein, the AI-informed epistemological perspective is “an order for a particular purpose, one out of many possible orders, not *the* order.”

In short, the contingency of AI-based conceptual metaphors resides in (1) its imposition of computational linguistic imagery and (2) the fact that this conceptual framing, as conceptual practice, fits the worldview of a certain in-group.

#### 4.2. The map-territory fallacy

Over time, the pervasive use of metaphors can lead to the risk of mistaking models (the metaphors) for reality (the actual system). This ‘map-territory’ fallacy occurs when the analogies (i.e., shared features between domains) are mistakenly regarded as constitutive of the system itself, promoting a conceptual equivalence with the model. As a result, the conceptual metaphor is epistemically stretched too far. *Contra* Favela (2022) – who states that “describing brains and minds as ‘computers’ is not metaphorical,” because “cognitive psychologists and AI researchers have literally meant that brains and minds are computational devices. That is to say, brains and minds are information processing systems in the same way that computers are” – we argue that the essentially metaphorical nature of the explanation is not always recognized. Whether or not it is *meant* literally is beside the point; the metaphorical conceptual organisation masquerades as ‘neutral’ scientific investigation or as inference to the best explanation. Like cybernetics, ‘information processing’ is a wide metaphysical framing, but *performing computations like a computer*, is a metaphorical subclass of that. The projection of the ontology of the metaphorical model into the system, as if it weren’t a contingent ‘net’, masks its metaphorical origin.

The conceptual metaphor, as a *set* of metaphors, constitutes analogical family resemblances between domains. One domain is modelled via another, guiding our understanding. Metaphorical models work when guiding our attention to high-level similarities between systems while obscuring others that are ill-fitting for the model (Möck, 2022), which is the template of abstraction via analogy: “the framing of the investigation of one order of nature by means of its similarity to a system whose order of relations are better known or more readily comprehended by the scientist” (Chirimuuta, 2020, p. 441). The many ways the systems differ are put to the side. Hence, for these metaphorical models to be epistemically fruitful, they have to abstract away the differences between the systems and emphasise their similarities.

Consider the ‘brain-computer’ metaphor. To comprehend brains and minds as computers requires understanding them as information processors. Conceiving of cognition like this serves as

fundamental guiding commitment of research in AI, cognitive psychology, and neuroscience (Favela, 2022). Such a conceptual framework had epistemic power because of its focus on high-level functions (e.g., input-output processes) while disregarding neurobiological details unshared by computers and brains (Chirimuuta, 2020). This framework excuses the otherwise unjustified separation between a system and its processes, allowing for a focus solely on the *function* of interest. The rest classifies as *just* metabolic support (Chirimuuta, 2024). Therefore, the computational metaphor of the brain is an abstracted model of cognition that focuses on the *function* rather than on the biological substrate supporting it.

Abstraction plays a fundamental role in science (Chirimuuta, 2020). For example, as van Rooij, et al. (2024) argue, it helps researchers build computational models of cognition that become essential for exploring, hypothesising, explaining, and ultimately understanding the target domain (Van Rooij, 2022). Conceptual metaphors work *because* they abstract away disanalogies and simplify complex concepts into more accessible forms, ultimately emphasising similarities while suppressing differences (Chirimuuta, 2020, 2024). Metaphors are abstracted linguistic models of phenomena.

We observe that metaphorical language is a useful tool for grasping epistemically opaque phenomena. But metaphors' epistemological guidance comes with risk of distracting or manipulating our epistemic access (Möck, 2022). Specifically, the danger lies precisely in metaphors' consistent suppression of differences and amplification of high-level similarities. Recall Kompa (2021) saying that metaphors exploit *contingent* rather than *defining* features. *Overemphasis* of one particular metaphorical model risks gradually eroding our awareness of the distinctions between source and target systems., to the point where "we forget that they ever existed" (Chirimuuta, 2020, p. 441). Thereby the risk of conceptual substitution increases. Conceptual substitution qualifies as gradually occurring conceptual elimination, without justification. This 'conceptual forgetfulness' has been a particularly persistent issue throughout the history of AI (Birhane, 2021). As AI systems increasingly mimic human-like behaviours, conceptual substitution occurs when AI systems are perceived as direct replacements for human cognition (Chirimuuta, 2020). Additionally, our limited understanding of intelligence and overly confident predictions about AI, reinforces the belief that AI concepts explicate essential features of cognition and that we thereby embody such concepts ourselves (Birhane, 2021; Chirimuuta, 2020). We cannot rule out *a priori* that there is an avenue for conceptual engineering, but that avenue needs our *justification*, not *habituation*.

Ignoring differences between machines and organisms risks extrapolating non-defining mechanical or algorithmic features to human behaviour and psychology. Such an outcome has profound epistemic implications. Among the epistemic risks is that conceptual replacement narrows our perspective of the target system and ends up limiting our understanding of it. Particularly, if we regard organisms only through the mechanical metaphor, we overlook aspects ill-fitted to this comparison, but they are fundamental to the system. We forget the organism is more than its functional replication in the machine we draw from to analogize (Chirimuuta, 2020). Moreover, this totalizing view reduces human cognition to *calculating machines* (Baria & Cross, 2021; Floridi &

Nobre, 2024; Fuchs, 2021). To speak with Kompa’s criteria for epistemic success: essential properties are obscured, not all inferential patterns can be preserved in the projection, and disanalogies persist.

Returning to the stochastic toddler example, over time we have come to see our cognition and humanity through the lens of AI/computers. The perspective that we are meat-made AI systems isn’t bizarre anymore; “i am a stochastic parrot, and so r u”. The phrasing implies that humans and AI are fundamentally the same in terms of their core processes: Human and machine language and cognition emerge from statistical computations, extrapolating from patterns learned from the training data supplied by perception, the unsupervised learning of neuronal networks, and the supervised learning of interaction with the world. This reductionist view neglects the unique aspects of human cognition, encouraging the mistaken belief that the similarities are all there is.

However, we shouldn’t detract from the epistemic benefit of the metaphor as high-level abstraction. There are similarities between human language-use and LLM language production. For example, both humans and AI systems’ linguistic activities are grounded in language that is shared, public and historical (Birhane & McGann, 2024). In that sense, we can understand language learning in humans *as if* they are LLMs. But even though these systems behave in human-like ways, their constitution is fundamentally different (Shanahan, 2024). While their linguistic ‘outputs’ can be made to converge, human and LLM language ‘input’ processes strongly diverge, as do the meanings of ‘shared’ and ‘historical’ for both systems. Language learning in humans involves much more than just linguistic exposure; a human infant is born into a community of language-users with which it shares a limited world of embodied practices and the acquisition of language involves the interaction with this community and the world they share, whereas a LLM is a disembodied computational entity that, after training on gigantic corpuses of textual data, is able to predict the next word in a sequence of words (tokens) (Shanahan, 2024). As Birhane & McGann (2024, p. 5) aptly emphasise, language is active, embodied and dynamic and involves “voice, text, gestures, body languages, tones, pauses, hesitations, as well as what has been left unsaid”. The dimensions of language LLMs can master are only a subset of what human language involves. In Anthony Chemero’s (2023) words: “Although humans are quite facile with and can learn quickly from text-based information (just as LLMs do), interacting with text is only one of our ways of knowing about the world around us”. Even though at times LLMs might be an illuminating metaphor for understanding certain aspects of human language-learning-and-production processes, it’s crucial to understand that *as metaphor* it’s an incomplete representation of human language-use. We should remain wary of confusing the map with the territory.

### 4.3. The double metaphor problem

Turning to the foundations of the concept of computation, we argue that underlying AI cases of MCO is the *double* conceptual metaphor ‘INTELLIGENCE is ARTIFICIAL’. As Floridi & Nobre (2024) noted, AI research depends on cognitive terms like memory, neuron, stimulus and learning. Similarly, Mitchell (2024a) writes:

The field of AI has always leaned heavily on metaphors. AI systems are called “agents” that have “knowledge” and “goals”; LLMs are “trained” by receiving “rewards”; “learn” in a “self-supervised” manner by “reading” vast amounts of human-generated text; and “reason” using a method called chain of “thought.” These, not to mention the most central terms of the field—*neural* networks, machine *learning*, and artificial *intelligence*—are analogies with human abilities and characteristics that remain quite different from their machine counterparts.

This marks the first layer of the double metaphor: after the initial metaphorical borrowing, technical usages within the AI domain were charged with new technical auras, while being put to new uses, like (un)supervised learning and neural networks.

The second layer of INTELLIGENCE is ARTIFICIAL is the metaphorical connection underlying the concept of computation itself. Alan Turing (1937) mechanised the mathematical concept of computation (the human computer). Turing used cognitive terms like ‘state of mind’, seeing and remembering, but “Turing machines [do not] demonstrate or possess cognitive abilities; on the contrary, Turing was to stress that ‘machine intelligence’ only emerges in the shift from ‘brute force’ to ‘learning’ programs” (Shanker, 1987, p. 626). His quest revolved only about proving equivalence in terms of expressive power. Turing proposed a mechanical alternative for the human computer, which computes *via* cognitive abilities like seeing, remembering what has been seen and altering what has been remembered according to arithmetical rules, without actually having them. However, Wittgenstein was already aware of the fallacious nature of drawing epistemological conclusions from Turing’s success in the mechanisation of calculation through this metaphorical connection. How? According to Shanker (1987, p. 619), Wittgenstein (1980) called Turing’s machines “*humans* who calculate”, because “Turing [had] actually defined human calculation in mechanical terms so as to license the application of quasi-cognitive terms to the operations of his machines”.<sup>12</sup>

Furthermore, Shanker shows that Wittgenstein pre-empted any epistemological conclusions by connecting mechanisation to rule-following. The inequivalence of regularity and rule-following hinges on the logical grammar of rule-following.<sup>13</sup> Turing machines belong to our shared logical fundament, because of the way we *use* recursive effectively calculable functions: they form a logico-grammatical certainty (Shanker, 1987, p. 624; Wittgenstein, 1975). Like certainties, rules fixate the use of concepts and so does the rule of computation that the Turing machine represents: *anything that can be sensibly called an algorithm is executable by a Turing machine*. But this doesn’t imply that Turing machines follow meaningless rules because their execution is *mechanical*.

---

<sup>12</sup> Cf. Birhane, 2021.

<sup>13</sup> Supposedly Turing machines ‘follow meaningless rules’ (i.e., algorithms), and, because they are a mechanical formalisation of human computation, this would imply that human computation is also guided by mechanical rules. The problem with this inference is the idea of a ‘meaningless rule’. This problem figures in Wittgenstein’s examples of a caveman producing regular sequences of signs and chimpanzees scratching regular figures into the earth: a rule can be constructed to “describe the regularity”, but that doesn’t mean the observed regularity reduces to following a rule (Shanker, 1987, pp. 620-621).



Rather, “mechanically following a rule” is itself problematic, because rule-following is essentially *normative* (Wittgenstein, 2009): it can be evaluated based on the pattern the act of following emulates. The apparent confusion lies in conceiving of algorithms as meaning nothing because of their reducibility to atomic operations. But Turing’s machines simply cannot *account for* whether or not a rule is used for “regulation of their conduct” or whether or not the results that the machines produced were correct (Shanker, 1987, p. 638). Rule-following – like a psychological concept such as reading – is completely independent of any *internal* mechanism (Wittgenstein, 2009, §157), like the ‘state of mind’ Turing’s machine operatively depends on. Sure, rules can be followed ‘in a mechanical fashion’, but the rule itself should be presentable as ground for doing so.<sup>14</sup> Thus the nexus Turing machine/rule-following is already *metaphorically* rather than metaphysically charged, because the “basic fallacy” of the cognitive framework surrounding Turing machines is “to move from the indisputable simplicity of the sub rules of an algorithm to the conclusion that such procedures are intrinsically mechanical” (Shanker, 1987, p. 641).

Connecting this second metaphorical layer to the first layer makes INTELLIGENCE is ARTIFICIAL *doubly* metaphorical. Making the originally cognitive and behavioural concepts (Turing machine as metaphorical model of human rule-following) explain the human domain again goes full-circle. Others have also noted this full-circle movement, but explain it differently. Barwich & Rodriguez (2024, p. 14) argue the intertwinement of brain and computer

...unfolded through a dynamic, bidirectional exchange that enhanced our understanding of each system by viewing each through the perspective of the other. Turing’s work serves as a prime example of such reciprocal exchange. His insights into computational learning, initially inspired by child development studies [...] subsequently informed psychological approaches to childhood education, illustrating a full-circle influence.

Such a “bidirectional exchange” isn’t epistemically infertile. However, the Wittgensteinian investigations lead us to emphasise that whatever terrain is mapped, be it neural, psychological or behavioural, metaphors might help clarify mechanical or functional aspects of an organism, but only by emphasising specific aspects of a complex phenomenon while *downplaying or ignoring* others. Metaphors offer a partial and simplified view of the complex realities they aim to represent. Because of these downsides, we shouldn’t employ AI/computer metaphors unreflectively – they have epistemic success once we suspend our disbelief in modelling human cognition, behaviour and psychology as more or less complex Turing machines. Successful models should surely attract the attention they deserve, but not without the disclaimer that, *contra* computationalism (Odell, 2001) other epistemological conceptual constellations are available and possible.

---

<sup>14</sup> As a sidenote, the rise of LLMs has meant a new wind blowing for the problem of aligning technology to human values. Interestingly, in Pérez-Escobar & Sarikaya (2024) a Wittgensteinian conception of rule-following is drawn upon in order to inspire how to address the alignment problem in LLMs: how to make them ‘act’ in accordance with rules like humans can.

But one catch remains: the novel semantic charge derived from the technical AI application could still provide cues or avenues for conceptual engineering because of the conceptual metaphor's *descriptive invariance*.

## 5. Avenues for computational conceptual engineering of the human domain

We contend that AI advancements offer a unique, albeit limited, opportunity for conceptual engineering, i.e., refinement of often ambiguous and fuzzy behavioural and psychological concepts. However, caution is advised to prevent oversimplifications. If any form of *CE* is warranted, this should be because this new conceptual organisation of intensions and extensions yields better results in terms of a population *Y*'s understanding of the phenomenon *X* in question.

### 5.1. Opportunities for AI-inspired conceptual engineering

Subsequently, we explore what INTELLIGENCE is ARTIFICIAL shows that could provoke conceptual amelioration, elimination or expansion.<sup>15</sup>

*Amelioration?* Firstly, it could be that the conceptual metaphor *extends* human psychological, behavioural and cognitive concepts: widening the domain of application of concepts like thinking, seeing, remembering, feeling to include AI. This *epistemic* amelioration presents anomalies like AI systems as new cases of experiential access. But this is problematic, because, contrary to the mereological fallacy that forgets to take into account behavioural criteria for concept application (Bennett & Hacker, 2003), *only* functionally behavioural criteria are met by AI systems. So it would be a functionalist temptation to infer that the underlying processes (constitution, environmental embedding) would be the same (Mollema, 2024b); the conceptual core that ties into the shared form of life is gone. So tweaking extension doesn't represent any epistemic refinement, because it would simultaneously obscure standing conceptual criteria.

Alternatively, the full-circle application of the computational conceptual domain could recalibrate the concept of intelligence's *intension* via *semantic* amelioration. The options at hand to do so are changing a concept's pragmatic coordination, logical demarcation, or moral charge. Now, do any of these apply?

We argue moral and pragmatic amelioration aren't triggered. Firstly, widening the circle of intelligence is an indicator for changing a moral stance towards a being (van Gulick, 2024), but so far, INTELLIGENCE is ARTIFICIAL offers no moral reason for changing the standing concept of intelligence. Secondly, for pragmatic amelioration, a practical social motivator is lacking; sure,

---

<sup>15</sup>Deroy (2023, p. 887) discusses three ways in which conceptual approaches to AI and humans can be connected. (1) the *extension view* holds that "naive users extend their category of humans to include AI, even though they do not view AI as central or prototypical in that category"; (2) the *novelty view* is that "naive users have a new category for AI that shares some features with the category of humans but is distinct"; and (3) the *semi-propositional* view, "which proposes that naive users hold a non-fully propositional set of beliefs about AI that do not form a consistent concept, reminiscent of elements of religious beliefs, such as ghosts or spirits". We are inspired by (1) and (2) in what follows.

we *could* come to different agreements about when to speak of intelligence, but changing intension is not the same as coordinating applicability to selective extensions.

So we are left with the alethic option of repartitioning intelligence's logical space. For example, the ties of intelligence to human forms of life could be loosened. If redefined as 'the capacity to build models', rather than as loose clustering of cognitive capacities ('the intellect'), intelligence's domain of application widens, not only to AI systems, but also to other animals, while partially preserving its core. But given the use-guided grammar of this logical space, this is the hardest to do *well*. A new definition is proposed to fit the new usages of AI-inspired intelligence in human domains *and* intelligence itself in AI domains. Alternatively, the standing concept could be tweaked in more subtle ways, such as stripping the criteria related to (likeness to) cognition in order to accommodate new phenomena, i.e., non-cognitive intelligence. So rearrangement of the criteria for intelligence are what the conceptual metaphor's success points at, if we think it points at *changing* the concept of intelligence at all.

*Expansion?* But maybe INTELLIGENCE is ARTIFICIAL doesn't show any problems or imperfections in standing concepts. A possibly less parsimonious way to follow the metaphor's lead is to coin *new terms for new concepts* by opting for lexical expansion. This interprets the re-application of the doubly metaphorically charged concepts as momentum for concept creation. Not the expression of parallels between human intelligence and AI, but the articulation of a conceptual subdomain of artificial and natural intelligence should drive the re-application then. By allowing this to be, say, 'model building' or 'problem solving', a new concept is introduced which expands out of and connects AI and intelligence, but with a different semantic load. However, Cappelen's (2020, p. 140-146) caveats for illegitimate lexical expansions should be overcome:

1. *preserving lexical effect*: existing terms can incite important moral, legal, cognitive or emotive effects that shouldn't be unnecessarily distorted by new concepts;
2. *topic continuity*: the worry that because words are used over time "say the same thing", keeping the underlying concepts stable is important to facilitate diachronical discourse;
3. *anchoring role*: if the parent concept plays an important role in clustering instances, a lexical expansion with it as starting point might not be such a good idea as it good distort this clustering role; and
4. *role in social ontology*: "changing the meaning of a lexical item might contribute to a change in social reality".

In short, while it sounds appealing to coin new terms for new forms of expression, it is harder to do than it seems, apart from the fact that for the take-up of such a concept, a need to embed it into a conceptual practice should be present (Queloz, 2021) such that its reasonably intended also becomes the pragmatic use.

*Elimination?* Given the requirement that some *Y*'s understanding of *X* needs to be unequivocally improved by conceptual engineering, we warn against practising elimination inspired by AI, because of the longstanding tradition of reducing logico-grammatical conceptual levels to underly-

ing mechanical concepts. The danger of opting for conceptual elimination in the guise of conceptual substitution is inspired by the reductionistic character of INTELLIGENCE is ARTIFICIAL, such as of ‘the LLM as symbol for human language’ or ‘the mind as a computer’. While reduction is always tempting, in the *CM*-case of AI it’s a one sided emphasis of the explanatory value of mechanico-computational perspectives at the cost of the messy organismic properties of life and a disregard for the “motley” (Queloz & Cueni, 2021) and “rough ground” (Wittgenstein, 2009, §107) of our conceptual practices. Contrary to spelling out any benefits of conceptual elimination, we advise opting instead for careful amelioration of existing concepts’ intensions or the warranted introduction of new concepts.

## 5.2. Two challenges for AI-inspired conceptual engineering

The main challenge is the tension between conceptual engineering and conceptual ethics. Not every engineering opportunity is also *desirable*<sup>16</sup> and it’s questionable for *whom* this conceptual engineering in terms of AI is useful. Within the AI domain, the ethical question whether it is desirable to widen extensions, change intensions, or eliminate a concept has to be answered. What does the conceptual change mean in terms of the ethical relations that morph along with it? For intelligence, change affects inclusion into the moral circle, and the extent to which anthropo-exceptionalism is possible by appealing to human’s superior intelligence. Furthermore, as the controversy between meat chauvinism and computational chauvinism at this paper’s start showed, some hedged positions on the AI/intelligence relationship are neck-deep into other societal and economic commitments, such as employment by BigTech. These commitments import non-philosophical presuppositions into the debate. An example of this is what Deroy (2023, )p. 888) calls “cultural match”: different pragmatic usages of “conceptual boundaries” render “official public discourse comprehensible or acceptable to citizens” with disregard for metaphysical boundaries. Alternatively put,, legal or cultural recalibration of concepts to accommodate AI can change the conceptual constellations that are tied to specific cultural forms of life of certain social groups and not others, without being necessarily grounded in *actual* differences between humans and AI. If we follow the lead of the conceptual metaphor, then, paraphrasing Rosenthal (1982, p. 294), language-users would be normalized towards a “metaphorical naming” that makes political things *invisible*. In other words, there is a danger inherent in solidifying this perspectival shift via an engineering move, as it “predispose[s] us in favor of a specific line of action and it is because metaphors embody proposals that they produce this effect” (Ankersmit, 1993, p. 162). Therefore conceptual engineering of this domain cannot be done without dealing with the related ethical question of *what* conceptual form will serve *whose* purposes. Conceptually hygienic engineering shouldn’t be driven by societal narratives and corporate framings of AI capabilities.

Moreover, while narrowly defined concepts of cognition may be desirable from a scientific perspective to study its constitutive parts, this view risks flattening the mind’s richness to purely

---

<sup>16</sup> Haslanger’s moral form of semantic amelioration is a typological example of this.

mechanistic and reproducible functions (Floridi & Nobre, 2024). The brain isn't merely an information-processing or computational apparatus, but a living, plastic, and dynamic system (Favela, 2022; Fuchs, 2021). Conceptual engineers must recognize that cognition and the brain are inherently complex systems. Ethically, this is crucial. Reducing cognition to purely computational or mechanistic terms not only impoverishes our understanding of the mind but also stimulates anthropomorphising of AI systems. This misrepresentation contributes to the ongoing phenomenon of AI hype, where capabilities and performance are overstated (Barrow, 2024; Placani, 2024). If the brain is likened to a computer, we risk also seeing computers as brains (Baria & Cross, 2021; Birkhane, 2021). This bidirectional computational metaphor has profound ethical implications, urging us to carefully consider the societal and philosophical consequences of the concepts we adopt.


These challenges don't settle the debate over the applicability of conceptual engineering to the AI-human nexus. We leave the articulation of further challenges to future work.

## 6. Conclusion

Are we all unsupervised learners, stochastically quibbling our way through life? Is intelligence essentially artificial and is the computational the 'best game in town' for explaining the human conceptual domain? By articulating the conceptual metaphor and conceptual engineering approaches to these questions, we showed that the conceptual metaphor that underlies these comparisons of the AI domain with the human domain is severely limited. These comparisons represent only a contingent epistemological perspective: the organisation of the human domain in terms of computational concepts, which abstracts away from the target domain into an emphasis of shared properties. Furthermore, we argued it also risks committing the 'map-territory fallacy': mistaking the model of the target system – built out of computational concepts – for the *actual* target system to be explained. Lastly, by turning to Wittgenstein's reaction to Turing's concept of computation, we explained the appeal of the conceptual metaphor rooted in AI by unmasking it as *doubly metaphorical*: firstly rooted in a metaphorical cognitive connection between the human computer and its mechanical formalisation as Turing machine, and secondly by the full-circle re-application of the concepts *after* being charged with new meanings from AI research. With respect to the alternative – the conceptual engineering view – we argued INTELLIGENCE is ARTIFICIAL points to avenues for alethic amelioration of intelligence's intension, or the introduction of a new concept for the shared non-cognitive capacity underlying AI and intelligence. To conclude, we pointed to two challenges for AI-inspired conceptual engineering.

Understanding human domains in terms of AI is yet another partially insightful, partially flawed instance of MCO. At its worst, "Seeing a living human being as an automaton is analogous to seeing one figure as a limiting case or variant of another; the cross-pieces of a window as a swastika, for example" (Wittgenstein, 2009, §420). At its best, it's one that prompts us to reflect anew on how the boundaries of our current concepts serve us and how they could be approved.

## Bibliography

- Aaronson, S. (2024). The Problem of Human Specialness in the Age of AI. *The Problem of Human Specialness in the Age of AI*. <https://scottaaronson.blog/?p=7784>
- Altman, S. (2022, April 12). *I am an stochastic parrot and so r u* [X]. <https://x.com/sama/status/1599471830255177728?lang=es>.
- Ankersmit, F. R. (1993). Metaphor in Political Theory. In *Knowledge and Language: Volume III: Metaphor and Knowledge* (Ankersmit, F. R. & Mooij, J. J. A. (eds.), pp. 155–202). Springer.
- 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* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2107.14042>
- Barrow, N. (2024). Anthropomorphism and AI hype. *AI and Ethics*, 4(3), 707–711. <https://doi.org/10.1007/s43681-024-00454-1>
- Barwich, A.-S., & Rodriguez, M. J. (2024). Rage against the what? The machine metaphor in biology. *Biology & Philosophy*, 39(4), 14. <https://doi.org/10.1007/s10539-024-09950-4>
- 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>
- Bennett, M., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Blackwell.
- Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, 27(1), 44–61. [https://doi.org/10.1162/artl\\_a\\_00336](https://doi.org/10.1162/artl_a_00336)
- Birhane, A., & McGann, M. (2024). Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences*, 106, 101672. <https://doi.org/10.1016/j.langsci.2024.101672>
- Black, M. (1955). Metaphor. In *Philosophical Perspectives on Metaphor* (Johnson, M. (ed.), pp. 63–82). University of Minnesota Press.
- Black, M. (1962). *Models and Metaphors*. Cornell University Press.
- Black, M. (1977). More About Metaphor. *Dialectica*, 31(3/4), 431–457.
- Boden, M. A. (1988). *Computer models of mind: Computational approaches in theoretical psychology*. Cambridge University Press.
- Brette, R. (2022). Brains as Computers: Metaphor, Analogy, Theory or Fact? *Frontiers in Ecology and Evolution*, 10, 878729. <https://doi.org/10.3389/fevo.2022.878729>
- Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press.
- Cappelen, H. (2020). Conceptual Engineering: The Master Argument. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (1st ed., pp. 132–151). Oxford University Press. <https://doi.org/10.1093/oso/9780198801856.003.0007>
- Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11), 1828–1829. <https://doi.org/10.1038/s41562-023-01723-5>

- Chirimuuta, M. (2020). The Reflex Machine and the Cybernetic Brain: The Critique of Abstraction and its Application to Computationalism. *Perspectives on Science*, 28(3), 421–457. [https://doi.org/10.1162/posc\\_a\\_00346](https://doi.org/10.1162/posc_a_00346)
- Chirimuuta, M. (2021). Your Brain Is Like a Computer: Function, Analogy, Simplification. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms* (Vol. 17, pp. 235–261). Springer International Publishing. [https://doi.org/10.1007/978-3-030-54092-0\\_11](https://doi.org/10.1007/978-3-030-54092-0_11)
- Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. The MIT Press.
- Churchland Smith, P. (1980). A perspective on mind-brain research. *Journal of Philosophy*, 77, 185–207.
- Cobb, M. (2020). *The Idea of the Brain: The Past and Future of Neuroscience*. Profile Books.
- Colombo, M., & Piccinini, G. (2024). *The Computational Theory of Mind* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781009183734>
- Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Allen Lane.
- Deroy, O. (2023). The Ethics of Terminology: Can We Use Human Terms to Describe AI? *Topoi*, 42(3), 881–889. <https://doi.org/10.1007/s11245-023-09934-1>
- Favela, L. (2022). Complexity: Understanding brains and minds on their own terms. *Preprint*. [https://www.researchgate.net/publication/358402375\\_Complexity\\_Understanding\\_brains\\_and\\_minds\\_on\\_their\\_own\\_terms](https://www.researchgate.net/publication/358402375_Complexity_Understanding_brains_and_minds_on_their_own_terms)
- Felin, T., & Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Decision Making. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4737265>
- Floridi, L., & Nobre, A. C. (2024). Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, 34(1), 5. <https://doi.org/10.1007/s11023-024-09670-4>
- Fuchs, T. (2021). Human and Artificial Intelligence: A Clarification. In T. Fuchs, *In Defence of the Human Being* (1st ed., pp. 13–48). Oxford University PressOxford. <https://doi.org/10.1093/oso/9780192898197.003.0002>
- Haslanger, S. (2020). Going On, Not in the Same Way. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (1st ed., pp. 230–260). Oxford University PressOxford. <https://doi.org/10.1093/oso/9780198801856.003.0012>
- Hopster, J., Brey, P., Klenk, M., Löhr, G., Marchiori, S., Lundgren, B., & Scharp, K. (2023). 6. Conceptual Disruption and the Ethics of Technology. In I. Van De Poel, L. E. Frank, J. Hermann, J. Hopster, D. Lenzi, S. Nyholm, B. Taebi, & E. Ziliotti (Eds.), *Ethics of Socially Disruptive Technologies* (1st ed., pp. 141–162). Open Book Publishers. <https://doi.org/10.11647/obp.0366.06>
- Hopster, J., & Löhr, G. (2023). Conceptual Engineering and Philosophy of Technology: Amelioration or Adaptation? *Philosophy & Technology*, 36(4), 70. <https://doi.org/10.1007/s13347-023-00670-3>
- Koch, S. (2023). Why Conceptual Engineers Should Not Worry About Topics. *Erkenntnis*, 88(5), 2123–2143. <https://doi.org/10.1007/s10670-021-00446-1>

- Kompa, N. (2021). Insight by Metaphor – The Epistemic Role of Metaphor in Science. In A. Heydenreich & K. Mecke (Eds.), *Physics and Literature* (pp. 23–48). De Gruyter. <https://doi.org/10.1515/9783110481112-002>
- Kövecses, Z. (2008). Conceptual metaphor theory: Some criticisms and alternative proposals. *Annual Review of Cognitive Linguistics*, 6, 168–184. <https://doi.org/10.1075/arcl.6.08kov>
- Kövecses, Z. (2017). Conceptual Metaphor Theory: Some New Proposals. *LaMiCuS*, 1(1), 16–32.
- Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 202–251). Cambridge University Press. <https://doi.org/10.1017/CBO9781139173865.013>
- Lakoff, G., & Johnson, M. (1981). Conceptual Metaphor in Everyday Language. In *Philosophical Perspectives on Metaphor* (Johnson, M. (ed.), pp. 286–325). University of Minnesota Press.
- Lakoff, G., & Johnson, M. (2003). *Metaphors We Live By*. University of Chicago Press.
- Lewontin, R. C. (1996). Evolution as Engineering. In J. Collado-Vides, B. Magasanik, & T. F. Smith (Eds.), *Integrative Approaches to Molecular Biology* (pp. 1–10). The MIT Press. <https://doi.org/10.7551/mitpress/3824.003.0002>
- Löhr, G. (2024). If conceptual engineering is a new method in the ethics of AI, what method is it exactly? *AI and Ethics*, 4(2), 575–585. <https://doi.org/10.1007/s43681-023-00295-4>
- Machine Learning Street Talk. (2024a). *Joscha Bach—Why Your Thoughts Aren't Yours*. YouTube. <https://www.youtube.com/watch?v=3MkJEGE9GRY>
- Machine Learning Street Talk. (2024b). *“We Are All Software”—Joscha Bach*. YouTube. [https://www.youtube.com/watch?v=34VOI\\_oo-qM](https://www.youtube.com/watch?v=34VOI_oo-qM)
- 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), 12. <https://doi.org/10.1609/aimag.v27i4.1904>
- Miller, E. F. (1979). Metaphor and Political Knowledge. *The American Political Science Review*, 73(1), 155–170.
- Mitchell, M. (2024a). The metaphors of artificial intelligence. *Science*, 386(6723), eadt6140. <https://doi.org/10.1126/science.adt6140>
- Mitchell, M. (2024b). The Turing Test and our shifting conceptions of intelligence. *Science*, 385(6710), eadq9356. <https://doi.org/10.1126/science.adq9356>
- Möck, L. A. (2022). Prediction Promises: Towards a Metaphorology of Artificial Intelligence. *Journal of Aesthetics and Phenomenology*, 9(2), 119–139. <https://doi.org/10.1080/20539320.2022.2143654>
- Mollema, W. J. T. (2024a). *On certainty*, Left Wittgensteinianism and conceptual change. *Theoria*, theo.12558. <https://doi.org/10.1111/theo.12558>
- Mollema, W. J. T. (2024b). Social AI and the Equation of Wittgenstein’s Language User with Calvino’s Literature Machine. *International Review of Literary Studies*, 6(1), 39–55.
- Neisser, U. (1967). *Cognitive psychology*. Appleton-Century-Crofts.



- Nicholson, D. J. (2013). Organisms≠Machines. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(4), 669–678. <https://doi.org/10.1016/j.shpsc.2013.05.014>
- Odell, S. J. (2001). Wittgenstein and Computationalism. *From the ALWS Archives: A Selection of Papers from the International Wittgenstein Symposia in Kirchberg Am Wechsel*. <https://wab.uib.no/ojs/index.php/agora-alws/article/view/2412>
- Pérez-Escobar, J. A., & Sarikaya, D. (2024). Philosophical Investigations into AI Alignment: A Wittgensteinian Framework. *Philosophy & Technology*, 37(3), 80. <https://doi.org/10.1007/s13347-024-00761-9>
- Piccinini, G. (2006). Computational explanation in neuroscience. *Synthese*, 153(3), 343–353. <https://doi.org/10.1007/s11229-006-9096-y>
- Piccinini, G. (2009). Computationalism in the Philosophy of Mind. *Philosophy Compass*, 4(3), 515–532. <https://doi.org/10.1111/j.1747-9991.2009.00215.x>
- Pitkin, H. P. (1972). *Wittgenstein and Justice*. University of California Press.
- Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*, 4(3), 691–698. <https://doi.org/10.1007/s43681-024-00419-4>
- Queloz, M. (2021). *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*. Oxford University Press.
- Queloz, M., & Cueni, D. (2021). Left Wittgensteinianism. *European Journal of Philosophy*, 29(4), 758–777. <https://doi.org/10.1111/ejop.12603>
- Rivadulla, A. (2006). `Metáforas y modelos en ciencia y filosofía. *Revista de Filosofía (Madrid)*, 31(2), 189–202.
- Rosenthal, D. C. (1982). Metaphors, Models, and Analogies in Social Science and Public Policy. *Political Behavior*, 4(3), 283–301.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Shagrir, O. (2006). Why we view the brain as a computer. *Synthese*, 153(3), 393–416. <https://doi.org/10.1007/s11229-006-9099-8>
- Shanahan, M. (2024). *Simulacra as conscious exotica*. arXiv. <https://arxiv.org/abs/2402.12422>
- Shanker, S. G. (1987). Wittgenstein versus Turing on the Nature of Church's Thesis. *Notre Dame Journal of Formal Logic*, 28(4), 615–649.
- Taylor, C., & Dewsbury, B. M. (2018). On the Problem and Promise of Metaphor Use in Science and Science Communication. *Journal of Microbiology & Biology Education*, 19(1), 19.1.40. <https://doi.org/10.1128/jmbe.v19i1.1538>
- Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>

- van Gulick, R. (2024). Consciousness. In *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition, Zalta, E. N. & Nodelman, U. (eds.)). <https://plato.stanford.edu/archives/win2022/entries/consciousness/>
- Van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1(3), 127–128. <https://doi.org/10.1038/s44159-022-00031-5>
- Van Rooij, I., Guest, O., Adolphi, F., De Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*. <https://doi.org/10.1007/s42113-024-00217-5>
- Veluwenkamp, H., Hopster, J., Köhler, S., & Löhr, G. (2024). Socially Disruptive Technologies and Conceptual Engineering. *Ethics and Information Technology*, 26(4), 65. <https://doi.org/10.1007/s10676-024-09804-3>
- Wang, A. (2023, November 13). Is My Toddler A Stochastic Parrot? *The New Yorker*. <https://www.newyorker.com/humor/sketchbook/is-my-toddler-a-stochastic-parrot>
- Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus* (Trans. Ogden, C. K.). Kegan Paul, Trench, Trubner & Co.
- Wittgenstein, L. (1975). *On Certainty* (Anscombe, G. E. M. & von Wright, G. H. (eds.). Trans. Paul, D.). Blackwell.
- Wittgenstein, L. (1980). *Remarks on the Philosophy of Psychology, Volume I* (Trans. Anscombe, G. E. M., Anscombe, G. E. M. and von Wright, G.H. (eds.)). Basil Blackwell.
- Wittgenstein, L. (2009). *Philosophical Investigations* (Trans. Anscombe, G. E. M., Hacker, P. M. S. & Schulte, J.). Wiley-Blackwell.