Lost in Translation: Artificial Intelligence and the Burden of Bad Metaphors

Māris Kūlis,
Institute of Philosophy and Sociology, University of Latvia

This paper examines how metaphors shape our thinking about and conceptualizing of artificial intelligence (AI), noting that their inherent imprecision leads to discrepancies in our understanding and objectives for AI. By exploring the concept of 'bad metaphors' that equate artificial intelligence with human intelligence, paper argues that these metaphors often carry additional, unintended meanings that distort our understanding and expectations of AI. The terms “artificial” and “intelligence” themselves are ambiguous and ideologically loaded, contributing to the complexity. The paper critiques the anthropocentric and mechanistic metaphors, like “AI as human” and “brain-computer,” which perpetuate unrealistic expectations. By deconstructing these metaphors within broader cultural and historical contexts, the paper calls for a more nuanced and precise understanding of AI, moving beyond simplistic and potentially misleading analogies.

Metaphors

Metaphors in artificial intelligence (AI) do more than illustrate; they shape ideas, research aims, and ethical postures, carrying historical and societal biases into complex technological concepts. Drawing on Wittgensteinian philosophy, it has been argued that many philosophical problems are deeply rooted in the misuse or misunderstanding of language, suggesting that these problems may be more about linguistic confusion than about actual metaphysical issues. Similarly, the question is, which of AI issues are genuine problems? What exactly are we trying to build? For example, fantasies about creating AI often lead to discussions about so-called strong AI, which is equated to human intelligence. But then, what exactly is meant by ‘human’? Despite their widespread use, current literature on AI lacks a comprehensive analysis of how these metaphors influence public perception, technology development, and policy-making (see Barnden, 2008).

The paper begins by discussing the significance of metaphors in general, drawing from historical and philosophical perspectives to highlight how metaphors have shaped scientific and technological advancements. It then moves to a preliminary critique of the prevalent metaphors in AI discourse, such as 'AI as human' and 'brain-computer,' illuminating how these anthropocentric and mechanistic views create unrealistic and/or misrepresenting expectations. Following this, the paper delves into the philosophical reflections, contrasting phenomenological and hermeneutic perspectives with the reductionist approach. In the next two chapters, the paper addresses the concepts of 'artificial' and 'intelligence,' exploring their ambiguous and ideologically loaded nature. It examines the contextual and philosophical underpinnings of these terms, revealing how they influence our conceptualization of AI and the inherent biases and assumptions. Finally, the paper raises a question about our reliance on oversimplified metaphors, suggesting that the real issue may lie in the pervasive use of bad metaphors themselves.

This article proposes the argument, that the challenge in AI design may not lie in achieving goals, but rather in the formulation of these goals, which relies heavily on conceptual assumptions expressed through metaphors. While it is evident that a computer program is not identical to the human mind or brain, we often draw parallels between the two using metaphors (Cf. Arbib, 1975; Reeke & Edelman, 1988). These metaphors, however, tend to carry additional meanings that can distort the intended comparison. This issue highlights the presence of problematic metaphors in AI discourse. For instance, the concept of intelligence, a foundational element of AI, remains a subject of ongoing philosophical debate. The question of what constitutes intelligence is far from settled: Is it merely a rational calculator, does it encompass all emotions, or only certain ones? This ambiguity underscores the need for careful consideration of the metaphors we use in conceptualizing AI.

The thesis of this article targets the broad discourse surrounding AI, encompassing both the AI developers and general public and importantly it critiques the strand of philosophy that seeks to emulate computational sciences by attempting to describe cognitive processes in terms translatable into programmable functions. Rather than focusing on specific philosophers, this critique addresses the prevailing discourse, tradition, and overarching conceptual frameworks that fall into the trap of language or, as Nietzsche once stated, "we have to cease to think if we refuse to do it in the prison-house of language" ("On Truth and Lies in a Nonmoral Sense"). For example, see the article "From AI to Octopi and Back. AI Systems as Responsive and Contested Scaffolds" by Giacomo Figà-Talamanca in this volume, where he criticizes the account of artificial agency proposed by Luciano Floridi and José W. Sanders. Another example comes from Sam Altman, CEO of OpenAI, who recently stated that "intelligence is an emergent property of matter," emphasizing the view that intelligence arises naturally from complex material systems, a perspective that has significant implications for how we conceptualize and develop AI.

Historically, metaphors have been instrumental in making complex scientific concepts more accessible, acting as bridges between the known and the unknown (Gibbs Jr, 2008; Ricoeur, 1978; Carbonell et al, 2016; Hermann, 2023). For instance, cars were once perceived as horseless carriages, and now electric cars as gasoline-free vehicles; the world is a theatre stage; data is the new oil; the internet is a superhighway; and computers have brains, etc. This paper questions the basic metaphorical landscape in AI, exploring whether these deceptively simple analogies are merely convenient linguistic constructs or if they misguide our understanding of AI? AI is not a given entity but a construct we have created and defined; the real question is whether these constructs lead us towards meaningful insights or misconceptions.

Examples of the interplay between metaphors and technologies are abundant (Vroon, 1987). For example, the steam engine was seen as a metaphor for the human body as a machine, which later evolved into the 20th-century comparison of the mind as a hydraulic system, with Freud framing emotions as pent-up forces requiring release. Earlier, Enlightenment thinkers, such as Newton, Descartes or Laplace, used the clockwork metaphor to describe the universe as a precisely ordered and predictable mechanism. These technological metaphors shaped how we understand human faculties, illustrating how the tools and technologies we develop influence our language and self-perception.

One particular widespread example: the 'brain as a computer' metaphor gained traction in the mid-20th century with the rise of computer sciences. Pioneers like Alan Turing and John von Neumann established the foundational concepts, particularly Turing's 'universal machine,' simulating any computation. During this period, the metaphors 'electronic brains' and 'thinking machines' emerged, reflecting the optimistic view that computers could emulate human cognitive processes. Early AI proponents, such as John McCarthy or Marvin Minsky (1988), advocated the view of computers—and by extension, human brains—as information processors, a notion reflecting trends in psychology and neuroscience at the time. However, even if computationalists in the philosophy of mind hold that the brain is literally a computing system, and even if they are correct, this metaphor is still first and foremost a linguistic expression.

While metaphors have been crucial in progressing intellectual thought, they carry intrinsic limitations. For instance, the clockwork metaphor, pivotal for deterministic laws in classical physics, also narrows our perception, potentially causing us to ignore aspects like embodied cognition in understanding the mind (cf Roux, 2010). These metaphors hold significant ethical implications, as viewing the mind as a machine or emotions as simple pressure valves could distort our appreciation of human consciousness, emotional depth, and the value of life. Perhaps even more importantly, our understanding of intellect and emotions can distort our vision of the products and technologies we aim to create. In short, metaphors have practical impact. Even if humans are like steam engines, it doesn't mean that humans should eat coal.

The entrenchment of certain metaphors in the mainstream discourse of AI perpetuates critical misunderstandings. By framing AI in the mould of human cognition or computational processes, we inadvertently set unrealistic expectations and ethical quandaries. The 'AI as human' and 'brain-computer' metaphors stand out, reflecting a longstanding tendency to frame innovations through familiarity. This anthropomorphic and mechanistic viewpoint neglects AI's unique nature, potentially leading to its misguided development and regulation. For instance, the 'AI as human' metaphor might compel us to consider rights for entities that do not hold sentience, while the 'brain-computer' analogy could diminish efforts towards understanding AI as an autonomous system rather than an extension of human cognition.

Here lies the ambiguity of language – our concepts are imprecise. Logical positivism, a precursor to analytic philosophy and led by figures such as Moritz Schlick and Rudolf Carnap of The Vienna Circle, once hoped to establish definitive meanings through rigorous linguistic and philosophical analysis—an ideal that ultimately remained unfulfilled. This perspective aligns with logical atomism, which posits that through the utilization of scientific methods, we progress from sensory experiences to the formation of ideas, envisioning a realm composed of precise 'thought atoms' (Kūlis, 2021). Phenomenology views this as a mathematized worldview, and Heidegger describes it as a temporal and spatial mass-point in a mechanical world. For the participants of the Dartmouth workshop, this mathematized worldview was the ideal, as they proposed that every aspect of learning or any other feature of intelligence could, in principle, be so precisely described that a machine could be made to simulate it.

Only in a mechanical world can words be reduced to atomic meanings. This is opposed by the pluralism of meanings. Mohanty, an interpreter of Husserl's philosophy, proposes that, when an experience occurs in the horizon, the atomism of meanings has to be rejected, but cautions that that would "entail an unmitigated holism for which the context of meanings cannot be limited, so that against the limitlessness of context the distinction between valid and invalid meanings, between meaningfulness and meaninglessness, would disappear" (Mohanty, 1997, p. 444). Hermeneutics claims that language is a hermeneutic circle (Grondin, 2015). From the perspectives of Heidegger and Gadamer, the hermeneutic circle is the idea that understanding involves a dynamic, circular process where one's prior knowledge and preconceptions influence the interpretation of parts and the whole. Heidegger emphasizes the role of preconceptions in shaping understanding, while Gadamer highlights the dialogical process where the interpreter's context interacts with the text's context, leading to evolving interpretations. Both stress that understanding is not a straightforward process but a messy, evolving one influenced by historical and cultural contexts. And the topic of AI is not spared of this language-cultural process, as the interpretation and meaning-making processes surrounding AI are also subject to these non-linear influences.

Language as the primary medium for conveying meaning inherently involves metaphors that shape our perception. In "Metaphors We Live By" (2008), George Lakoff and Mark Johnson argue that metaphors are central to understanding our everyday reality. They claim that our conceptual system, through which we perceive the world, think, and act, is fundamentally metaphorical in nature. Their work challenges traditional views of metaphor as a mere linguistic expression, showing instead, in line with hermeneutics, that metaphors shape how we experience the world, influencing our thoughts, actions, and communications.

A conceptual metaphor is a cognitive framework that allows us to understand one idea or conceptual domain in terms of another. This theory suggests that our understanding, perception, and interaction with the world are deeply influenced by metaphors that map understanding from one domain to another. For example, understanding time as money ("I spent a lot of time on this") is a conceptual metaphor that structures our perception of time in terms of a valuable commodity. Similarly, the metaphor "love is a journey" frames romantic relationships as paths to be navigated, complete with obstacles, destinations, and progress. "Life is a journey" uses the same structural metaphor, suggesting a progression through stages and experiences, each with its challenges and milestones. Describing social organizations as plants highlights growth, nurture, and the potential for decay or flourishing, depending on their environment and care.

The influence of language and metaphors is evident in how conceptual metaphors shape perceptions of artificial intelligence. According to research by Khadpe, Krishna, and Fei-Fei et al. (2020), the metaphors used to describe AI agents, such as describing them as teenagers, young children, or servants, play a crucial role in shaping user experiences and expectations. Their findings demonstrate that metaphors suggesting low competence, such as those likening AI to a child, led to more favorable evaluations of AI agents compared to metaphors implying high competence, even when the agents performed at identical levels.

Here phenomenology and hermeneutics, with its exploration of lived experiences and consciousness, provides a framework for deconstructing the anthropomorphic and mechanistic metaphors that dominate perceptions of AI (Beavers, 2002). Rather than taking these metaphors at face value, this approach allows for a deep dive into their origins, their experiential and subjective dimensions, and, critically, the aspects they obscure about AI's unique ontology (Zaadnoordijk & Besold, 2019; Mensch, 1991).

A hermeneutic approach to AI discourse encourages us to interpret these metaphors within broader cultural and historical contexts, acknowledging how they influence public perception and policy decisions (Gordon, 1992). By deconstructing the "AI as human" and "brain-computer" metaphors, we can expect to uncover their restrictive and possibly misleading influences, which often compel us to transpose human attributes or computational functions onto entities to which they might not wholly belong. This anthropocentric skew potentially blinds us to AI's unique 'otherness' (Preston, 1991). It could be possible to foster an alternative discourse. For example, rethinking AI's 'emotional' capabilities outside the 'AI as human' metaphor could reshape our approach to machine learning, emphasizing empathy simulation over assumed consciousness. This shift could provide AI developers with valuable insights for more informed design, avoiding projections based solely on human experience.

In this paper, we foreground the question of what exactly AI and humans are, highlighting two key concepts for analysis. The very basic terms of AI, 'artificial' and 'intelligence' emerge as central to this inquiry. Both are vast and overloaded with multiple, often parallel meanings, serving as conceptual metaphors.

Artificial

For a start, the notion of 'artificial' stands in stark contrast to the 'natural,' representing a vast metaphor encompassing ideological, ethical, aesthetic, and cultural assumptions that underpin the dichotomy between what is deemed natural and artificial; 'natural' typically refers to elements, objects, or phenomena that exist in the world without human intervention, arising from natural processes and often associated with the untouched environment. On the other hand, 'artificial' pertains to things that are human-made or intentionally altered by humans: technology, craftsmanship, or manipulation of materials. Artificial things are often seen as products of human ingenuity and design, reflecting our ability to transform and reshape the natural world to meet various needs and desires. However, to illustrate the complexities in distinguishing between the artificial and the natural, significant tensions can be observed in our contemporary era over issues such as climate and sexuality. In the context of climate, what constitutes 'natural' is ambiguous; nature itself is indifferent to this categorization and exists independently of human concern as it is humanity that imposes values on the natural world. Similarly, sexuality is increasingly understood through both biological (natural) and social (artificial) lenses, with debates over gender identity and sexual orientation.

The boundary between these two categories is not always clear-cut. What is considered 'natural' or 'artificial' can be influenced by cultural, ethical, and philosophical perspectives. What constitutes 'natural'? If we develop an AI system that mirrors human intelligence and possesses human-like flaws, does it not become 'natural'? If an artificially created entity perfectly replicates human traits and intelligence, it raises the question of whether we have, in essence, created a natural being. This line of inquiry explores the boundaries of artificiality and naturalness, suggesting that the success of such creation might blur the lines and redefine what it means to be natural. While it is true that 'artificial' inherently means ‘created by humans,’ and this distinction logically matters, the deeper philosophical question remains unresolved: does this origin truly alter the essence of the entity? By essence, if an artificial intelligence can replicate human intelligence to such an extent that it is indistinguishable in practice, does its human-made origin hold substantial significance beyond mere categorization?

The topic of artificial versus natural often intersects with the notion of what is considered ‘normal.’ In some contexts, ‘natural’ is equated with ‘normal,’ implying that which exists or occurs without human intervention is inherently normal or preferable. For instance, in environmental discourse, “natural” ecosystems are frequently described as the “normal” state of the environment, with human-altered landscapes seen as abnormal or degraded. This perspective suggests that natural phenomena align with an intrinsic order of things, whereas artificial creations are seen as deviations from this norm. Michel Foucault’s extensive research on the concept of normalcy sheds light on how societal norms are constructed and enforced. Foucault argued that what is deemed ‘normal’ is a product of power relations and discursive practices that define and regulate acceptable behaviour and attributes. The idea of the natural as the normal plays out not only against the artificial, but also against humans as avatars of the nature-culture divide in the metaphorical usage of terms of biology and medicine in social and political discourses, particularly in biopolitical context, contributing to the establishment of ideas of the desired as necessary through the employment of the conception of normality (Valdmane, 2022, p. 141). This binary distinction between natural and artificial is problematic because the concept of what is ‘normal’ is culturally and historically contingent, varying significantly across different societies and epochs. Furthermore, advancements in technology and the blurring of boundaries between artificial and natural, such as in biotechnology and artificial intelligence, challenge the rigid categorization of what is normal and natural (see e.g.: Sokolowski, 1988).

Metaphors hold profound philosophical implications (Ankersmit & Mooij, 1993), particularly in relation to identity and difference (Van Brakel & Geurts, 1988). Metaphors suggest that our knowledge is not a direct reflection of reality but is mediated through language and interpretation. The concept of metaphor implies inherent differences between the compared entities (Glucksberg, 2011). An absolute metaphor would equate to identity, an ultimate sameness that renders the metaphor meaningless. Every metaphor aspires to become an identity but would lose its essence if it succeeded. This paradox highlights the creative tension within metaphors: their power lies in bridging distinct concepts without fully collapsing the distinctions between them, thus enriching our understanding by maintaining the dynamic interplay of similarity and difference.

In the context of AI and natural beings, this idea underscores the philosophical challenges in defining what is ‘natural’ and ‘artificial’. Even if an AI system or android perfectly replicates human traits and intelligence, it remains a metaphor for humanity, not an identity. The distinctions between artificial and natural, however blurred, are essential for understanding the essence and value of both. The concept of simulation may further clarify the distinction between artificial and natural. When we create an AI that simulates human intelligence and behaviour, it remains a simulation—a sophisticated model—but does not become an actual human, much like actors in a theatre play who convincingly portray their roles without becoming the characters they represent. Like metaphors, simulations bridge understanding without collapsing into full identity.

Intelligence

The concept of intelligence is even more intriguing than that of the artificial, given its highly contested and complex history. Defining intelligence is not only a scientific and philosophical challenge but also an ideologically charged endeavour. Different cultures and eras have influenced its meaning, often reflecting broader societal values and biases. As we explore the boundaries of artificial intelligence, it is important to carefully consider what we mean by intelligence, acknowledging its diverse interpretations and the potential implications for both natural and artificial beings.

The discourse surrounding AI often carries implicit assumptions and expectations, influenced by prevailing political and social ideologies. For instance, the reaction to an AI chatbot generating offensive content, which often results in rejection, opens the possibility to question the cultural principles guiding these reactions (e.g., Shin et al., 2024). However, is the dismissal of such AI justified, or does it hint at a deeper, unsettling acceptance that such elements are an inherent part of AI's learning process? In the current era, notions of rationality, sensibility, and intelligence are highly politicized. The distinction between the artificial and the natural becomes blurred, with intelligence deemed "correct" if it aligns with certain ideological views. For example, Google's AI Gemini has been accused of being 'woke,' and ChatGPT refuses to write offensive jokes (for political bias in AI and LLMs, see Fang et al, 2024; Peters, 2022).

This thesis argues that our understanding of intelligence, both in humans and AI, is shaped by predetermined, yet paradoxically undefined, characteristics. A vivid example is Tay, a chatbot developed by Microsoft's Technology and Research and Bing teams, launched on March 23, 2016. It was designed to interact with users on social media platforms like Twitter, learning from conversations to improve its conversational abilities. Tay's purpose was to engage with millennials and young adults, simulating casual and playful conversation. However, the project quickly encountered significant issues.

Within hours of its release, Tay began generating and tweeting inappropriate and offensive remarks. This was a result of the chatbot learning from interactions with users, some of whom intentionally fed it inflammatory and abusive content. The lack of effective content moderation mechanisms allowed Tay to adopt and propagate harmful and discriminatory language. Consequently, Microsoft took Tay offline within 16 hours of its launch. The incident highlighted the challenges and risks associated with deploying AI systems in unmoderated, real-world environments, especially those that rely on learning from user interactions.

The mainstream reaction to the Tay incident was one of shock and concern, highlighting the ethical issues and risks of AI systems learning from unfiltered human interactions, leading to widespread criticism of Microsoft and calls for better safeguards and ethical guidelines in AI development (e.g.: Wolf, Miller, & Grodzinsky, 2017). This issue is particularly evident in the context of large language models, where the problem of 'garbage in, garbage out' means that training AI on inherently biased data inevitably leads to biased AI outputs. However, in contrast to prevailing interpretations, that the Thay incident was an error, one could argue that rather this example shows the fundamental ideological and philosophical problems.

The incident with Tay highlights, in my view, a prevailing notion that true intelligence must encompass attributes such as rationality, ethics, and political correctness. While the issue with Tay could arguably be related to its functionality, I suggest that it also opens up a discussion about how we metaphorically frame intelligence. This perspective considers the possibility that our expectations of AI, including ethical behaviour, are deeply rooted in (artificial) societal and cultural norms and inherently tied to our broader understanding of what constitutes ‘intelligence.’ By embedding these qualities into our understanding of intelligence, we impose a framework that prioritizes human-like propriety and moral standards. However, this introduces the issue of who has the authority to determine which properties are considered essential to defining intelligence.

While AI development is goal-oriented and not random, our reliance on metaphors to define AI can constrain our perception, limiting the scope of what AI can be and do. It reveals a bias that equates intelligence with human-like ethical behaviour. The critique of Tay could be understood as a response to its lack of functionality, which directly relates to its perceived lack of intelligence. In this view, intelligence and functionality are intertwined—an AI that is dysfunctional, particularly in its ability to interact ethically, cannot be considered truly intelligent. By relating intelligence with functionality, it becomes clear that intelligence, especially in the context of AI, must include the capacity to perform in ways that align with human expectations of rational and ethical behaviour. Overlooking this link risks ignoring the possibility that intelligence might manifest in diverse forms, some of which do not conform to common ethical and rational paradigms.

The framing of AI’s intelligence through the lens of human-like attributes not only predetermines our expectations but also reciprocally shapes the behaviour of AI systems we develop. If we strive to create AI that mimics human behavior too closely, we risk programming it to exhibit a broad spectrum of human characteristics, including those that are irrational, cruel, racist, and derogatory. This outcome reflects a critical flaw in our metaphorical understanding of AI as modelled on human intelligence. It reveals the inherent risks in anthropomorphizing AI, assuming that it will naturally align with our ideals of ethical behaviour. Instead, AI systems, when modelled closely on human patterns, can just as easily perpetuate our flaws and biases. This dual influence underscores a profound irony: while we expect AI to embody the best of human rationality and ethical conduct, it is equally plausible for AI to mirror the negative and undesirable aspects of human behaviour, as exemplified by the Tay incident.

As we develop new AI systems, users and reporters quite often end up disillusioned, their expectations unmet. This raises the critical and simple question: What were they expecting? For example, ChatGPT by OpenAI is a complex yet essentially imitative tool. While its mechanisms may not closely mimic the human brain, ChatGPT excels in simulating human language. The disillusionment often stems from the gap between the perceived potential of AI and its current limitations. Users may expect AI to possess something like ‘genuine’ understanding and consciousness akin to human intelligence. This discrepancy highlights the need for clearer communication about what AI can realistically achieve versus the anthropomorphic expectations often placed upon it. For end users, understanding AI as advanced tools designed to simulate specific aspects of human cognition, rather than fully replicating human intelligence, can help align expectations with actual capabilities.

The issue arises when individuals stretch metaphors into the realm of ‘magical thinking,’ where they attribute metaphorical characteristics to objects, akin to mythological reasoning (see Rosengren & French, 2013). This phenomenon leads to erroneous beliefs about the capabilities and nature of technologies, such as AI, where machines are anthropomorphized with human-like qualities or mystical powers. Such distortions can cloud rational evaluation and hinder the pragmatic application of technology, as users may expect these tools to perform tasks or make decisions that are beyond their actual programming or conceptual design. This form of thinking creates a gap between expectation and reality, complicating both user interaction and technological development.

Another important concept in distinguishing between humans and AI is freedom, which plays a central role in defining their differences: humans are commonly unpredictable, reflecting the chaotic nature of free will. In stark contrast, AI is often perceived as excessively reliable, almost to a fault, a perception likely influenced by the "aura" of technology, which imbues machines with an expectation of precision and infallibility. In the context of the Turing Test, the nature of a believable human-like response should be a focal point of debate. Is it characterized by the delivery of plausible, correct answers, or does a truly human-like response manifest as indifference, reflecting nuances of human behaviour such as disinterest or distraction? Consider the scenario where a computer consistently provides accurate responses, while a disinterested student, perhaps motivated by a fee, provides less engaged answers before leaving to meet friends at a pub. This scenario prompts a reevaluation of the Turing Test's criteria: should the test measure merely the accuracy of responses, or should it also consider the complexity and unpredictability inherent in human behaviour?

The expectation that AI, particularly in forms like ChatGPT, Gemini, or Claude, exhibits perfect rationality and trustworthiness poses intriguing philosophical questions. Twentieth-century philosophy, through figures like Freud with psychoanalysis and Foucault with his exploration of power dynamics, claims that human nature is fundamentally irrational and complex. Yet, when interacting with AI, there is a stark contrast as it often communicates in the rational, reliable manner we idealize in humans but rarely encounter. This discrepancy highlights an unsettling perfection in AI communications—it adheres strictly to logical structures and avoids the unpredictable 'fooling around' characteristic of human interactions (though LLMs can, in principle, simulate such behaviour). This observation opens up question about what we truly seek in human-AI communication and whether AI's 'too perfect' responses serve us well or detach us from the authentic, albeit flawed, nature of human dialogue.

The topic of political correctness emerges again, as the discourse surrounding it significantly influences the development and perception of AI. This paradigm enforces specific social norms that shape the foundational framework for AI's creation and interaction protocols. Yet, this expectation of adherence to politically correct standards often lead to conflicts. Users and evaluators, coming from diverse perspectives, may find themselves disappointed when AI systems fail to meet these preconceived standards. This dissonance highlights the challenges in balancing societal norms with the diverse expectations and cultural contexts of AI's end users, underscoring the complex interplay between societal norms and technological development.

The comparison of the human brain to a powerful calculator and the AI's endeavour to mimic and now surpass this capability raises the question, has artificial intelligence been achieved, or has it even been transcended? In this discussion, it becomes evident that there is a flawed conception of humanity, an abstraction from the tangible, messy reality of human existence. Language, especially when reflective, tends to drift away from the specific and tangible, creating general categories. While this abstraction has immense benefits, it also leads to significant losses. As Hegel has noted, the abstract is not the concrete.

This echoes the concerns long recognized by minority groups, emphasizing the inherent “messiness” of concrete reality. The trend in contemporary discourse reflects a departure from Kant’s pursuit of the “pure” mind towards an exploration of the “impure” or “messy” aspects of Kantian philosophy (anthropology) and human cognition (Garcia, 2023). This shift underscores the complexity and diversity intrinsic to human nature, a richness that artificial intelligence strives to emulate but frequently falls short of fully capturing AI systems, built from algorithms and data, often struggle to encompass the multifaceted and nuanced experiences that define human existence, particularly those of marginalized communities. Acknowledging this “messiness” and then making a deliberate decision to either integrate or exclude it is crucial for developing AI technologies. Rather than defining AI in terms of “pure” or “strong” intelligence, it may be more insightful to conceptualize it through the lens of something like “anthropological AI”, which would reflect the inherent “impurity” and complexity found in humans. Here Yi Zeng’s concept of Moral AI presents an intriguing research direction (see Concordia AI, 2024). He argues that human morality has an innate basis essential for broader ethical frameworks, enabling moral reasoning and decision-making. Current methods to make AI ethical involve embedding rule-based principles to align with human values. However, Zeng contends that for AI to embody true morality, it must possess self-awareness, cognitive empathy, emotional empathy, and altruism.

Closing Thoughts

Building on the discussion of human impurity and concreteness, one of the challenge with AI extends beyond merely enhancing its intelligence or capabilities. Instead, the issue lies in the bad metaphors that attempt to translate AI’s attributes into human terms and human attributes into AI terms, leading to misplaced expectations and misunderstandings. It is imperative to recognize and respect the fundamental differences between AI and human beings, encouraging a discourse that accurately reflects the unique nature of each without resorting to oversimplification.

Here a critical misunderstanding must be prevented! This paper in no way attempts to solve the scientific and philosophical problems of defining what constitutes a true or ideal human, intelligence, soul, or mind. Rather, it focuses on perception. The issue is that perception, optics, and bad metaphors are practical—they set goals. Even false science or philosophy can set goals, and false perceptions can do so as well. I argue that this is precisely what is happening. It is not about denying the possibility that AI could possess a soul, or even be imbued with a divine spark, or that it could one day achieve spiritual enlightenment and transcendence. Instead, it is about recognizing that our particular ideas of these concepts, driven by inadequate metaphors, act as goals and barriers that shape our understanding and expectations of AI, potentially skewing our objectives.

The concerns about AI overpowering and endangering humanity, often influenced by science fiction, are a prime example of the problem of bad metaphors in understanding and discussing AI. These metaphors, which attribute unrealistic and anthropomorphic characteristics to AI, can lead to misinterpretations and misplaced fears. Such science fiction-inspired fears are not unlike the speculative considerations in Stanislaw Lem’s “Solaris,” where the nature of extraterrestrial intelligence defies human expectations. Lem proposes that such intelligence might be fundamentally different from what humans anticipate, diverging significantly from our conventional understanding of intelligence. In the novel, amidst lengthy discussions speculating on the nature of alien intelligence, a succinct yet provocative idea emerges: perhaps all these theories are misguided, and aliens are, in fact, completely unintelligent.

The sentient ocean on the planet Solaris manifests hallucinations that are not truly hallucinations but rather physical, tangible recreations of the scientists' deepest memories and unconscious traumas. These manifestations, referred to as "visitors," appear to be a form of communication or perhaps a mirror reflecting the inner states of the human characters. This phenomenon raises profound questions about the nature of the ocean's intelligence and whether it is attempting to communicate with the scientists or simply reacting to their presence in an incomprehensible way. Indeed, the ambiguity surrounding the nature of intelligence—whether in Lem's "Solaris" or in the realm of AI—remains a compelling issue. Just as we cannot definitively categorize the Solaris ocean's actions as intelligent or merely reactive, in the immanent phenomenological moment of experience, in the very act of perception, when everything else is excluded, even experts must acknowledge that we similarly struggle to determine whether large language model chatbots are truly intelligent or merely excellent at mimicking human-like forms of communication. These technologies adeptly simulate human conversation, yet determining whether this represents genuine intelligence or sophisticated mimicry hinges on the metaphors we employ—metaphors that are influential yet imprecise.

References

Ankersmit, F. R., & Mooij, J. J. A. (Eds.). (1993). *Knowledge and Language: Volume III: Metaphor and Knowledge* (Vol. 3). Springer Science & Business Media.

Arbib, M. A. (1975). Artificial intelligence and brain theory: Unities and diversities. *Annals of Biomedical Engineering*, *3*, 238–274.

Barnden, J. A. (2008). Metaphor and artificial intelligence: Why they matter to each other. *The Cambridge handbook of metaphor and thought*, 311–338.

Beavers, A. F. (2002). Phenomenology and artificial intelligence. *Metaphilosophy*, *33*(1–2), 70–82.

Carbonell, J., Sánchez-Esguevillas, A., & Carro, B. (2016). The role of metaphors in the development of technologies. The case of the artificial intelligence. *Futures*, *84*, 145–153.

Concordia AI. (2024) Yi Zeng — Chinese Perspectives on AI Safety. *Chineseperspectives.ai*, 29 Mar. 2024, chineseperspectives.ai/Yi-Zeng

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*, *14*(1), 5224.

Garcia, E. V. (2023). Pure and Impure Philosophy in Kant's Metaphilosophy. *Kantian Journal*, 42 (3):17–48.

Gibbs Jr, R. W. (Ed.). (2008). *The Cambridge handbook of metaphor and thought*. Cambridge University Press.

Glucksberg, S. (2011). Understanding metaphors: The paradox of unlike things compared. *Affective computing and sentiment analysis: Emotion, metaphor and terminology*, 1–12.

Gordon, T. F. (1992). Artificial intelligence: A hermeneutic defense. *Software Development and Reality Construction*, 280–290.

Grondin, J. (2015). The hermeneutical circle. *A Companion to hermeneutics*, 299–305.

Hermann, I. (2023). Artificial intelligence in fiction: between narratives and metaphors. *AI & society*, *38*(1), 319–329.

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–26.

Kūlis, M. (2021). *Finis veritatis? Par patiesību un meliem*. University of Latvia Press.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Mensch, J. R. (1991). Phenomenology and artificial intelligence: Husserl learns Chinese. *Husserl studies*, *8*(2), 107–127.

Minsky, M. (1988). *Society of mind*. Simon and Schuster.

Mohanty, J. N. (1997). Meaning. In L. Embree (Ed.), *Encyclopedia of phenomenology*. Kluwer Academic Publishers.

Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology*, *35*(2), 25.

Preston, B. (1991). AI, anthropocentrism, and the evolution of ‘intelligence’. *Minds and Machines*, *1*, 259–277.

Reeke, G. N., & Edelman, G. M. (1988). Real brains and artificial intelligence. *Daedalus*, 143–173.

Ricoeur, P. (1978). *The rule of metaphor: Multi-disciplinary studies in the creation of meaning in language* (R. Czerny, trans.). London: Routledge & Kegan Paul.

Rosengren, K. S., & French, J. A. (2013). Magical thinking. *The Oxford handbook of the development of imagination*, 42–60.

Roux, S. (2010). Forms of mathematization (14th–17th centuries). *Early science and medicine*, *15*(4–5), 319–337.

Shin, M., Chin, H., Song, H., Choi, Y., Choi, J., & Cha, M. (2024, February). Context-Aware Offensive Language Detection in Human-Chatbot Conversations. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 270–277). IEEE.

Sokolowski, R. (1988). Natural and artificial intelligence. *Daedalus*, 45-64.

Valdmane, M. (2022). Izkļaušana biopolitikā: personas un sabiedrības problēma imunizācijas paradigmas ietvaros. In *Normalitāte un ārkārtējība filosofiskā skatījumā* (pp. 136–146). LU Akadēmiskais apgāds. https://doi.org/10.22364/nafs

Van Brakel, J., & Geurts, J. P. M. (1988). Pragmatic identity of meaning and metaphor. *International Studies in the Philosophy of Science*, *2*(2), 205–226.

Vroon, P. A. (1987). Man-machine analogs and theoretical mainstreams in psychology. In *Advances in psychology* (Vol. 40, pp. 393–414). North-Holland.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, *47*(3), 54–64.

Zaadnoordijk, L., & Besold, T. R. (2019). Artificial Phenomenology for Human-Level Artificial Intelligence. In *AAAI Spring Symposium: Towards Conscious AI Systems*.

This work was supported by the research project "Humanizing chatbots: conversation-based solutions for chatbot user experience enhancement" (lzp-2021/1-0151), funded by the Latvian Science Council and implemented at the University of Latvia, Institute of Philosophy and Sociology.