Artificial and Human Intelligence: Algorithmic Bias and the Myth of Neutrality

di Seydel Ducosquel

This article explores the differences between artificial and human intelligence, focusing on algorithmic bias and epistemological neutrality in AI systems. It argues that AI, while capable of complex tasks, lacks true understanding and adaptability inherent in human intelligence. The article discusses challenges like algorithmic bias, the limitations of AI's knowledge processing, and the difficulty of achieving epistemological neutrality. It concludes that a multidisciplinary approach is essential to address these biases and their societal impacts.

Keywords: Artificial Intelligence; Human Intelligence; Algorithmic Bias; Epistemological Neutrality; Machine Learning

When talking about the definition of AI, the term is often used ambiguously, leading to misinterpretations and conceptual confusion. Can AI truly *know* or *understand*, or is it merely executing programmed tasks? While Natural Language Processors (NLP) may produce results that appear similar to human responses, these systems operate statistically manipulating vast databases rather than engaging in authentic cognitive processes. This highlights the crucial distinction between data processing and knowledge generation. This is often obscured with the use of comparisons with the human brain and other similar analogies.

On the other hand, most studies based only in one form of the humanities on the topic of AI tend to lack strong ontological analysis of the phenomena, leading to conceptual weaknesses. Hence why a multidisciplinary approach, integrating both computer science, informatics and philosophy, is essential for a more accurate perspective. This means focusing on the functional aspects of AI, such as Machine Learning (ML) algorithms and data processing, rather than human-like processes. Recognizing AI's inherent complexity, we must strive for clarity in understanding its mechanisms, despite the ongoing technical and philosophical challenges.

1. AI vs. HI: The problem of defining intelligence beyond analogies

The question of AI's intelligence becomes particularly significant when it prompts us to reflect on the nature of our own intelligence. When we're asking about the intelligence of the machine, deep down we're questioning our understanding of our own intelligence. The main differences between human intelligence and AI lies in the origin of the intelligence (natural vs. man-made), but also in subjectivity (subject experience vs. algorithm *objectivity*); flexibility (capacity to adapt and adjust to the environment); creativity (wide creative solutions vs. limited innovation); and autonomy (independent decision making vs. limited autonomous decision-making).

This specially highlights the remarkable complexity and adaptability of humans, as well as the ongoing challenges faced by the field of AI in trying to replicate the full depth of human intelligence. At the core of the distinction between AI and human intelligence lies the architectural differences of their "computational frameworks". AI systems, like most traditional computing systems, are typically built upon von Neumann architectures, where information is processed sequentially through a Central Processing Unit (CPU) and stored in memory locations. This mostly rigid and deterministic structure contrasts sharply with the highly interconnected organic networks that we encounter in the human system.

For AI systems, what we call *knowledge* is encoded in the form of logical rules and data structures, and this is another key distinction, the matter of representation and information processing. The language that AI uses has its roots in binary code and ASCII, as these form the foundational layers upon which digital systems operate. Andriy Burkov, while talking about ML, says that what a typical "learning machine" does, is finding a mathematical formula, which, when applied to a collection of inputs (called "training data"), produces the desired outputs[1]. This means that for the machine to simulate something, the input data must be always translated into binary code. The software that runs an AI can *understand* and process information only in the form of binary data and this means that all the input data (text, images, sensor readings, etc.) is diminished and converted into a numerical form and encoded into binary digits. In the same manner, the output provided by the AI is also in binary form, which is then converted into a "human-readable" format[2]. This translation goes through programming languages (Python, C++, Java, Lisp, etc.) and AI models (Neural Networks, Linear and Logistic Regression, Decision Trees, etc.)[3].

The topic of learning and adapting then comes into question. AI systems, particularly those based on ML techniques (supervised or unsupervised), learn from data, and improve their performance over time either in a supervised or non-supervised manner. Although, this *learning* is often confined to the specific task or domain for which the system was trained, with limited ability to transfer knowledge or skills to other contexts. The capacity for wide learning and application of knowledge is a pivoting point of human intelligence, and this cannot yet be replicated, despite the latest achievements of Deep Learning (DL)[4].

When talking about the definition of AI from a philosophical perspective there are usually two main ideas: weak AI (the idea that machines could act *as if* they were intelligent) and strong or general AI (the assumption that machines which act as if they're thinking, *are actually doing so*)[5]. But the matter at hand is far more complex.

In 1984 renowned computer scientist Edsger Dijkstra said «the question of whether a computer can think is no more interesting than the question of whether

1 - Cf. A. Burkov, *The Hundred-Page Machine Learning Book*, Open Access 2019.

2 - Cf. S. Russell - P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey 2011[3].

3 - Cf. M. Sipser, *Introduction to the Theory of Computation*, Cengage Learning, Boston 2012[3].

4 - Cf. Y. LeCun - Y. Bengio - G. Hinton (eds.), *Deep Learning,* in «Nature» 521, (2015) n. 7553, pp. 436-444.

5 - S. Russell - P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey 2011[3], p. 1035.

a submarine can swim»[6]. From this perspective, asking if AI can think would be a non-question, since machines have their own unique ways of operating that are fundamentally different from biological entities. The Turing test, designed in 1950 analyzes this topic, not asking if machines can think, but rather if they can pass a behavioral intelligence test[7]. However, this test, raises significant philosophical questions. When we state *it behaves then it is*, we fall in a basic confusion between ontology and epistemology, focusing on the *as if* (what we know about the thing) instead of the actual/possible existence of thought (the ontological part of it). This shift from understanding AI's potential *thought* to merely assessing its behavior reflects a fundamental confusion between what we know about AI and what AI might actually be.

Trying to apply one of the many perspectives on intelligence found in philosophy, we find that Spanish thinker Xavier Zubiri sees it as «not just a faculty for acquiring knowledge, but a way of apprehending reality»[8]. This means that, for there to be real intelligence, there needs to be a grounding in the environment that results in a grasping (apprehension) of said environment in an adequate way[9]. Zubiri was critical of the idea to reduce intelligence to simple information processing, since intelligence is not merely an abstract, disembodied process but is deeply rooted in sensory engagement with the world.

If we continue the idea of embodiment, we find British philosopher Andy Clark that says: «brains are first and foremost the control systems for biological bodies. Their primary function is to guide action and behavior in the real-world environments in which those bodies are situated»[10]. We can't consider the brain separately from the biological body, since cognition takes place within an environment. Therefore, for there to be true human intelligence we must see the system as a whole.

One prominent argument that is relevant to this debate, comes from John Searle's "Chinese Room" thought experiment. This experiment's final statement is that

---

6 - E. Dijkstra, *Selected Writings on Computing: A personal perspective*, Springer-Verlag, New York 1982, p. 17.

7 - A. Turing, *Computing Machinery and Intelligence*, «Mind» 59 (1950), n. 236, pp. 433-460.

8 - X. Zubiri, *Intelligence and Logos*, University Press of America, Washington DC 1983, p. 3.

9 - Of course, Zubiri is not the only thinker to follow this perspective. The basis of this idea comes from Aristoteles' adequation theory of truth, and another important concept that can be directly applied is the phenomenology and embodied cognition of Maurice Merleau-Ponty, that generally defines intelligence as the ability to engage with the world through perception and action in a dynamic and reciprocal relationship. Cf. M. Merleau-Ponty, *Phenomenology of perception*, Routledge, London 2012.

10 - A. Clark, *Supersizing the Mind: Embodiment, Action and Cognitive Extension*, Oxford University Press, Oxford 2008, p. 3.

mental states cannot be duplicated just on the bases of some program having the same functional structure with the same input-output behavior, rather than we would require that the program be running on an architecture with the same power as neurons. Searle maintains that there is something specific about the biological brain that can create real understanding, which cannot only be replicated in a simply physical symbol manipulation system[11].

When we apply this to AI it means that running the right program does not generate understanding or following Searle: AI systems are semantically blind[12]. Hence, the human complex knowledge experience cannot be translated into bits, or simply put: how can an AI process consciousness or subjective experiences? This argument is particularly used to remark the symbol-grounding problem[13] present in the Computational Theory of the Mind (CTM)[14].

This directly ties into the debate over whether AI can ever possess consciousness or subjective experience, as is evident in the unresolved question of whether Qualia can be encoded in computational systems. How can computation account for the *feel* of seeing red or feeling pain?[15] The specific way that something is experienced, the subjective, qualitative aspects of conscious experience, is still today a mayor challenge not only for technology, but for scientific knowledge in general. There is no consensus among philosophers or cognitive scientists on whether Qualia can be adequately translated into binary code; with some arguing that its possible in principle[16], while others maintain that the subjective nature of consciousness resists all reductive computational representations[17]. These issues

11 - Cf. J. Searle, *Minds, Brains, and Programs,* in «Behavioral and Brain Sciences» 3 (1980), pp. 417-424.

12 - «The computer has no way of attaching any meaning or semantic content to the formal symbols it manipulates. It is, in this sense, semantically blind» (J. Searle, *The Rediscovery of the Mind*, MIT Press, Cambridge, MA 1992, p. 207).

13 - The symbol-grounding problem questions whether abstract symbols can ever have meaning without a grounding in the lived experience of the agent.

14 - The CTM argues that the human mind operates like a computer, processing information through symbolic manipulation. Therefore, mental states (beliefs, desires, thoughts) are computational states, being the brain the hardware and the mind the software.

15 - For an in-depth analysis on Qualia: E. Wright (ed.), *The case for Qualia*, MIT Press, Cambridge 2008.

16 - The computational theory of mind holds that the human mind can be understood as a computational system, with mental states and processes corresponding to the manipulation of symbolic representations.

17 - This perspective is defended by those who state there is a «hard problem of consciousness», like David Chalmers that argues that the first-person nature of Qualia may be fundamentally irreducible to third-person objective descriptions like binary code.

however can influence the design, development and regulation of AI algorithm creation, especially considering the matter of bias and possible discrimination outcomes.

2. The Dilemmas of AI: Algorithmic Bias and Epistemological Neutrality

Algorithmic bias is defined as «systematic errors introduced into the outputs of a machine learning algorithm due to the way the algorithm is designed, the data used to train it, or the context in which it is developed»[18]. This specific bias (not the only bias in AI process) can manifest in the form of discriminatory outputs, skewed resource allocation, or the perpetuation of historical inequities. The source of algorithmic bias can be traced to various stages of the AI development lifecycle, from data collection and preprocessing to model training and deployment. The first issue is training data, specifically, its quality and this comes directly from the methods and protocols used to collect it. This means that the choice of objective functions, hyperparameter settings, or the inclusion (and/or exclusion) of certain features can amplify existing societal biases or even introduce new ones. The *truth* embedded in the algorithm's output is not objective since it reflects the biases inherent in the data it receives as input.

When we speak of algorithm bias, we are essentially discussing a mismatch between the algorithm's understanding of the world and the actual world. This discrepancy is not a mere error in programming, but a reflection of the ontological assumptions embedded within the algorithm. If an algorithm is trained on data that is skewed, it learns a biased representation of reality. This bias is not merely a superficial flaw but a deep-seated ontological issue. It questions the very nature of the reality the algorithm is constructing.

To mitigate the sources of algorithmic bias a multifaceted approach is needed, that considers technical, organizational, and policy-level interventions. Diversifying data collection and curation processes to ensure representative and inclusive datasets is vital as a first step. There has also been some intention to incorporate *algorithmic fairness* and debiasing techniques into the model design and training stages; this however can be difficult to achieve from a technical perspective, mainly because of the difficulty to access good and accurate data.

When we delve deeper into this topic another issue comes into question: the matter of epistemic fairness or neutrality in AI systems. Epistemological neutrali-

18 - H. Suresh - J.V. Guttag, *A framework for understanding unintended consequences of Machine Learning*, in «arXiv», http://arxiv.org/abs/1901.10002 (2019).

ty[19] is rooted in the philosophical principles of objectivity and impartiality. It is based on the idea that knowledge and decision-making should be based on empirical evidence, logical reasoning, and a commitment to truth, rather than being shaped by the personal beliefs, assumptions, or agendas of the knowledge producers or decision-makers[20]. This, however, does not seem to be an achievable goal for AI systems, mainly because, in this context, epistemological neutrality requires that the underlying algorithms, data sources, and training processes be designed and implemented in a manner that minimizes the influence of the developers' subjective perspectives and biases. Furthermore, «(a) system is neutral only if values external to the system's aim play a role in how the system delivers its results»[21].

To make the problem even more complex, many modern AI systems, particularly those based on Deep Learning, exhibit a degree of "opacity", making it challenging to understand the internal reasoning and decision-making processes within the AI itself. The *Black Box*[22] problem can obscure the influence of the developers' subjective perspectives and hinder efforts to ensure epistemological neutrality. The inherent opacity of AI models and the influence of various factors on their decision-making limit the potential to design AI systems that are completely neutral and free of biases.

However, from another perspective, with careful design, transparency and the integration of human oversight, AI could potentially overcome the limitations that are linked with human-centric epistemic practices; for this we require a multiple

19 - Epistemic neutrality refers to the «idea that a given epistemic agent, such as a scientist or a researcher, should strive to be impartial and unbiased in their pursuit of knowledge, setting aside personal preferences, ideological commitments, or other extraneous factors that could unduly influence their epistemic judgments» (H. Longino, *Science as social knowledge: Values and objectivity in scientific inquiry*, Princeton University Press, Princeton, NJ 1990, p. 62). From a more ontological perspective, epistemic neutrality however refers to »the view that the nature of reality, or the ontological structure of the world, is independent of and unaffected by the epistemic practices, beliefs, or conceptual frameworks of human knowers» (J. Searle, *The construction of Social Reality*, Free Press, New York 1995, p. 150).

20 - For further information on scientific objectivity: T. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago 1970[2].

21 - M. Phillips-Brown, *Algorithmic neutrality,* in «arXiv», http://arxiv.org/abd/2303.05103v3 last modified June 2024.

22 - Here it's important to note that the Black Box nature of informatic systems is an inherent aspect of their complexity. This opacity is not necessarily problematic; rather, it reflects the intricate layers of abstraction that define modern computing. The Black Box issue challenges our desire for transparency and control, but it does not inherently denote failure. Many systems function effectively without full interpretability, suggesting that the Black Box is a nuanced limitation rather than a fundamental flaw in informatics.

approach[23]. The first step would be to ensure the use of representative, inclusive, and debiased datasets that can help mitigate the influence of historical biases and provide a more balanced foundation for AI model training. Developing AI systems with built-in fairness constraints and enhanced explain-ability (like the case of XAI) can help reveal the underlying reasoning and reduce the impact of developers' subjective biases. Assembling AI development teams with a range of expertise, including philosophers, social scientists, and representatives from diverse backgrounds, can help surface and address epistemological biases.

Nevertheless, from a more critical perspective we are tempted to make an argument in another possible direction. Upon close examination, it becomes evident that the concept of epistemological neutrality is inherently human and its application to AI systems can be perceived as fundamentally flawed. Epistemology is deeply rooted in the human experience, the way we acquire, process, and validate knowledge is deeply linked to our cognitive structure, sensory capabilities, and the sociocultural contexts in which we operate. In contrast, AI systems, regardless of their level of sophistication, do not possess the same epistemological foundations as the human mind, nor do they exhibit the capacity for subjective, value-laden reasoning that is central to human knowledge acquisition. They do not possess the same sensory inputs, emotional responses, or lived experiences that shape human understanding of the world.

While AI systems can mimic problem-solving, they remain confined within the boundaries of their training data and the algorithms shaped by human intentionality. Their so-called *decisions* are mere logical deductions of pre-set parameters, lacking the autonomous epistemic agency found in human reasoning. AI's goal-oriented nature reveals an ontological limitation; it does not possess the self-reflective capacity to transcend its embedded biases or achieve epistemological neutrality. Rather than engaging in self-directed inquiry (which can be the first step to self-consciousness), AI functions as an extension of the designer's intent, perpetually bound to a mechanistic pursuit of predetermined objectives.

Though technical aspects of AI often minimize the significance of bias, its societal impact is undeniable and far-reaching. While AI is often regarded as a tool for objective data processing, it paradoxically has the potential to entrench existing stereotypes and perpetuate systemic discrimination. The ontological implications of algorithmic bias are profound because by learning from skewed data, the algo-

23 - Here we can mention the efforts in the development of Value Sensitive Design (VSD) which «offers a concrete methodology for how to intentionally embed desired values into new technologies. It consists of three iterative phases, namely conceptual/philosophical, empirical, and technical investigations» (J. Simon - P.H. Wong - G. Rieder, *Algorithm bias and the Value Sensitive Design approach*, in «Internet Policy Review» 9 [2020], n.4 http://doi.org/10.14763/2020.4.1534).

rithms not only produce flawed outputs, but actively shape the realities that are constructed through said outputs. Achieving epistemic fairness in AI systems demands an interdisciplinary approach that combines philosophical insights with practical, actionable strategies. Solely relying on philosophical critique risks overemphasizing abstract concerns without offering meaningful solutions. This limited approach might lead us to miss the crucial nuances needed to effectively mitigate bias and address the complex interactions between AI and society.

SEYDEL DUCOSQUEL
Post-doctoral Fellow in Philosophy at Sophia University Institute (Italy)
seydelducosquel@gmail.com