Izu: Journal of Critical Perspectives

ISSN E-3093-0650 (Online)
Vol. 4· No. 2· 2026
pp. 43–51

ARTICLE HISTORY
Received 27 Aug 2026
Revised 21 Sept 2026
Accepted 26 Oct 2026
Published 28 Oct 2026

KEYWORDS
Metaphysics
Artificial Intelligence
Mind
Functionalism

FUNDING
This research received no external funding.

LICENCE
© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

HOW TO CITE:
Alo, F. C. (2026). Metaphysical Assumptions as the Foundation for Artificial Intelligence. *Izu: Journal of Critical Perspectives, 4*(2), 43–51. https://doi.org/10.68251/IZU.4.1.6

RESEARCH ARTICLE https://doi.org/10.68251/IZU.4.1.6

Metaphysical Assumptions as the Foundation for Artificial Intelligence

Franklin Chibuzor Alo
Department of Philosophy, Federal University, Wukari
Correspondence: Chibuzorfrank208@gmail.com

ABSTRACT

Artificial Intelligence is often presented as a purely technical and mathematical enterprise. Yet beneath algorithms, data, and artificial neural networks lie deep metaphysical commitments about mind, matter and causation. This paper argues that AI systems are not metaphysically neutral. They presuppose specific answers to classical philosophical questions: For instance, what is intelligence? Can it be separated from embodiment? Is the mind reducible to computation? By tracing the historical and conceptual links between metaphysics and AI, the paper argues that Cartesian dualism, functionalism, and materialism shape AI's design and goals. Understanding these hidden assumptions is crucial for evaluating AI's capabilities, limits, and social impacts. The paper concludes that responsible AI development requires explicit engagement with metaphysics and not neutral nor avoidance of it.

Introduction

Metaphysics, traditionally regarded as the "first philosophy," is the branch of philosophy concerned with the ultimate nature of reality, being, existence, causality, identity, mind, and the fundamental structure of the universe. Rather than investigating particular objects, metaphysics asks what kinds of things ultimately exist and what it means for anything to exist at all. Every scientific discipline, whether consciously or unconsciously, rests upon metaphysical assumptions concerning the nature of reality and the intelligibility of the world since metaphysics is radically conceived as an idea. Artificial Intelligence, on the other hand, is not merely the science of constructing intelligent machines. At its deepest level, AI represents a philosophical attempt to understand whether intelligence, reasoning, learning, consciousness, and agency can exist independently of the biological human organism. Consequently, AI is not simply an engineering project but also a metaphysical project because its possibility depends upon prior assumptions concerning the nature of mind, intelligence, causation, and reality itself. Most importantly, no single thinker could be rightly said to invent the metaphysical basis of AI but rather a chain of ideas from ancient to early modern philosophers whose thoughts, directly and indirectly shaped the conceptual and its technical possibilities. Obviously, AI strongly relies on formalizing intelligence as rule-based manipulation of symbols, computation, and in some thinking, mechanistic processes. This manifests clearly in Aristotle's formal logic and rational deduction, Thomas Hobbes's reasoning as computation-mechanistic materialism, Descartes mind-body dualism and the thinking subject, as well as, Leibniz's calculable truth. In this paper, I shall attempt to demonstrate how the metaphysical assumptions of these great philosophers laid the foundation for artificial intelligence, and in extension, refutes the claims that philosophy does not bake bread or that it is a worthless discipline one shouldn't waste time studying.

Cartesian Rationalism and the Birth of the Thinking Machine

The philosophical lineage of artificial intelligence begins, perhaps surprisingly, with René Descartes in the seventeenth century. In his *Meditations on First Philosophy* (1641) and *Discourse on the Method* (1637), Descartes articulated a radical distinction between two fundamentally different kinds of substance: *res cogitans* (thinking substance or mind) and *res extensa* (extended substance or matter). This dualism, while philosophically contentious, had a lasting and productive consequence: it made it conceivable that thinking could be analyzed independently of the physical medium in which it occurs (Descartes, 1641/1998). Descartes famously argued that the essence of the self, the *cogito*, is purely a thinking thing. The body is a machine governed by mechanical laws, while the mind is a non-material substance whose defining attribute is thought. Although Descartes was skeptical that machines could truly think, he argued in the *Discourse on the Method* that animals and automata could never use language or reason in context-sensitive ways (Descartes, 1637/1998). His very framing of the problem established the conceptual terrain on which AI would eventually be constructed. By separating thinking from its biological substrate in principle, Descartes opened the door to the possibility that thought might be realizable in non-biological systems (Descartes, 1641/1998).

Margaret Boden, in her landmark work, *The Creative Mind: Myths and Mechanisms* (2004), traces how Cartesian mechanism prepared the ground for the cybernetic and computational thinking of the twentieth century. She argues that once mind was understood as a kind of process, however immaterial Descartes himself believed it to be, subsequent thinkers could progressively naturalize and formalize that process. The trajectory from Descartes' mechanical metaphysics to Leibniz's *calculus ratiocinator*, and from there to Boole's algebra of logic and eventually to Turing's universal machine, reflects a continuous deepening of the metaphysical assumption that thought is, at its core, a formal and rule-governed operation (Boden, 2004, pp. 45–50; Turing, 1950). The metaphysical assumption at stake here is what we might call the formalizability thesis: the belief that mental operations can be captured in precise, finite, explicit rules. This assumption is not empirically derived; it is a philosophical commitment about the essential nature of reason. It draws not only on Descartes but on the broader rationalist tradition, including Leibniz's dream of a universal language of thought and Kant's transcendental logic, and it provides the deepest metaphysical foundation for the very possibility of AI (Descartes, 1641/1998; Boden, 2004). More importantly, Hume treats the human mind as building complex thought from simple sensory impressions via three principles of association, namely: resemblance, contiguity, cause and effect. Hume made the foundational claim that intelligent behavior can emerge from associating experiential elements according to general, experience-driven laws. That is, the precise principle that the deep-learning systems operationalize.

Functionalism and Computationalism: The Philosophical Backbone of AI

The most directly influential metaphysical framework for artificial intelligence is functionalism, particularly in its computational form. Functionalism, developed most prominently by Hilary Putnam in the 1960s in works such as *Minds and Machines* (1960) and *The Nature of Mental States* (1967), holds that mental states are defined not by their intrinsic physical composition but by their functional roles, by the causal relationships they bear to inputs, outputs, and other mental states (Putnam, 1960, 1975). According to functionalism, what makes something a belief, a desire, or a pain is not the kind of physical stuff it is made of, but the role it plays in the broader cognitive economy of the system. This metaphysical commitment has profound implications. If mental states are defined by functional roles rather than physical constitution, then there is in principle no reason why a suitably organized silicon system could not have genuine mental states (Putnam, 1975). The mind becomes, in effect, a kind of software that can run on different physical hardware. This is the doctrine of multiple realizability: the same mental state can be instantiated in different physical substrates, carbon-based neurons or silicon-based transistors, so long as the functional organization is preserved (Putnam, 1975). Alan Turing's landmark paper, *Computing Machinery and Intelligence* (1950), instantiates precisely this functionalist metaphysics. Turing bypassed the question of what thinking "really is" by proposing a behavioral criterion: a machine thinks if it can perform indistinguishably from a human thinker in a conversational test, the now-famous Turing Test. Implicit in this proposal is the functionalist assumption that intelligence is constituted by what a system can do, not by what it is made of. Turing thus operationalized the functionalist metaphysics, transforming a philosophical thesis into a research program.

Computationalism extends functionalism by specifying the relevant kind of functional organization: it is computation. The mind, on the computationalist view, is a Turing-computable function over symbolic representations. This is the "physical symbol system hypothesis" articulated by Allen Newell and Herbert A. Simon in their influential paper *Computer Science as Empirical Inquiry: Symbols and Search* (1976). Newell

and Simon argued that a physical symbol system has the necessary and sufficient means for general intelligent action. This hypothesis is not an empirical generalization but a metaphysical conjecture about the nature of intelligence, a conjecture that has driven mainstream AI research for decades. Jerry Fodor's *The Language of Thought* (1975) gave the computationalist picture its most systematic philosophical articulation. Fodor proposed that cognition involves the manipulation of mental representations in a language-like format, the "language of thought" or "Mentalese," according to syntactic rules. This framework entails a particular metaphysics of mind: mental states are constituted by their formal, syntactic properties, and cognition is essentially a form of symbol manipulation. The computational metaphor is not merely a convenient tool for cognitive scientists; it embodies a substantive metaphysical claim about what minds are.

Physicalism, Materialism, and the Ontology of AI

While Cartesian dualism provided the initial conceptual opening for thinking about AI, the dominant metaphysical framework that has actually guided AI research is physicalism, the view that everything that exists is physical or supervenes on the physical. Contemporary AI research operates within a broadly physicalist ontology: the brain is a physical system, mental processes are physical processes, and understanding intelligence means understanding the physical mechanisms that implement it (Dennett, 1991). Physicalism as a background metaphysical assumption shapes AI in important ways. It rules out any appeal to non-physical entities or forces, souls, vital spirits, and immaterial minds as explanations of intelligent behavior. It demands that accounts of intelligence be reductive or at least naturalizing: intelligence must be explicable in terms of the natural sciences. This commitment is articulated clearly by Daniel Dennett in his *Consciousness Explained* (1991), where he argues that consciousness and intentionality, the two features most resistant to physicalist explanation, can be fully accounted for in computational and functional terms without remainder.

Dennett's "heterophenomenology" and his theory of the "intentional stance" which stands to build a fully materialist third-person account of the mind without invoking an unexplainable private inner world. These are efforts to preserve the explanatory utility of mental-state language within a fully physicalist ontology. By adopting the intentional stance, treating a system as if it had beliefs, desires, and rational agency, we can predict and explain behavior without committing to any non-physical ontology. This stance is applicable to thermostats, chess computers, and human beings alike, and it reflects the metaphysical assumption that mental properties are real but entirely explicable in physical and functional terms (Dennett, 1987). The physicalist assumption also underlies the enthusiasm for neural networks and deep learning in contemporary AI. The theoretical inspiration for artificial neural networks, from the early perceptrons of Frank Rosenblatt (1958) to the deep convolutional networks of Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, is that intelligence can be approximated by learning in physical systems that model the causal structure of the brain (Rosenblatt, 1958; LeCun et al., 2015). As Patricia Churchland argues in *Neurophilosophy* (1986), understanding the mind requires close attention to the physical structure of the nervous system, not just its abstract functional organization. Contemporary deep learning, as currently manifesting in AI, can be seen as an effort to cash out this physicalist intuition computationally.

Epistemological Assumptions: Knowledge, Representation and Rationality

AI is not only built on metaphysical assumptions about the nature of mind and matter; it also embeds specific epistemological assumptions about the nature of knowledge, the structure of rational inference, and the relationship between mind and world. These epistemological commitments are, in effect, applied metaphysics: they determine what counts as knowledge in AI systems and how that knowledge is represented and used (Haugeland, 1985). The classical AI tradition, symbolic AI or "Good Old-Fashioned AI" (GOFAI), as John Haugeland (1985) called it, was based on the epistemological assumption that knowledge can be explicitly represented as a set of propositions or rules. This is the doctrine of propositional or declarative knowledge: knowledge is essentially linguistic or symbolic in character, and an intelligent system must have access to a store of such representations. This assumption is Cartesian in spirit: just as Descartes believed that the mind's relationship to the world is mediated by clear and distinct ideas, symbolic AI assumed that an AI's relationship to its environment is mediated by symbolic representations (Haugeland, 1985, pp. 112–125).

The knowledge representation problem in AI, the challenge of encoding human knowledge in a form that computers can use, is fundamentally an epistemological problem, but it rests on the metaphysical assumption that knowledge has a determinate, formalizable structure. Projects like the Cyc knowledge base, initiated by Douglas Lenat in the 1980s, attempted to encode millions of common sense facts and rules in

formal logic, a Herculean effort that tacitly assumed that human knowledge is, in principle, fully articulable in explicit propositional form (Lenat & Guha, 1990). This assumption is contested by philosophers in the tradition of Michael Polanyi, who in *The Tacit Dimension* (1966) argued that much of human knowledge is tacit, known in practice but not explicable in explicit rules. The limits of GOFAI can partly be traced to this metaphysical tension. The epistemological assumption of rational agency, the idea that intelligent behavior can be modeled as expected utility maximization under uncertainty, underlies the decision theoretic and game theoretic approaches to AI. This assumption, rooted in the Bayesian interpretation of probability and the expected utility theory of John von Neumann and Oskar Morgenstern, treats the rational agent as an entity whose behavior can be fully characterized by its beliefs (probability distributions over states of the world) and its desires (a utility function) (von Neumann & Morgenstern, 1953). The metaphysical picture that underpins this framework is one in which rationality is a formal, context-independent property of the agent's internal representations, a view contested by bounded rationality theorists such as Herbert A. Simon and by situated cognition researchers who emphasize the role of the environment in shaping intelligent action (Simon, 1955).

The Phenomenological Critique: Dreyfus and the Limits of Formalism

The most sustained and philosophically sophisticated critique of the metaphysical assumptions underlying AI comes from the phenomenological tradition, particularly from the work of Hubert L. Dreyfus. In his landmark book, *What Computers Can't Do: A Critique of Artificial Reason* (1972), later revised and expanded as *What Computers Still Can't Do* (1992), Dreyfus mounted a systematic attack on the metaphysical foundations of classical AI, drawing on the phenomenological philosophy of Martin Heidegger and Maurice Merleau-Ponty (Dreyfus, 1972, 1992). Dreyfus identified four key metaphysical assumptions embedded in classical AI: the biological assumption (that the brain operates digitally), the psychological assumption (that the mind operates by manipulating discrete, context-free elements), the epistemological assumption (that all knowledge can be formalized), and the ontological assumption (that reality consists of independent, context-free facts). Each of these assumptions, Dreyfus argued, is contestable, and each reflects a specifically Cartesian picture of mind and world that phenomenology had already subjected to devastating critique (Dreyfus, 1972, 1992). Drawing on Heidegger's *Being and Time* (1927), Dreyfus argued that human beings are not primarily knowledge representing systems but beings in the world, entities whose relationship to their environment is characterized by practical engagement, bodily skillfulness, and what Heidegger called "readiness to hand." (Heidegger, 1962). The hammer is not first encountered as an object with a set of properties; it is encountered as something to hammer with. This practical, pre theoretical engagement with the world is not mediated by explicit representations; it is a form of knowing how, not knowing that. AI's exclusive focus on propositional knowledge, knowing that, thus reflects a fundamental misunderstanding of the structure of human intelligence.

Merleau-Ponty's *Phenomenology of Perception* (1945) deepened this critique by emphasizing the constitutive role of the body in cognition. For Merleau-Ponty, the body is not merely a vehicle for a disembodied mind; it is the very ground of perception and action. Skilled performance, whether playing the piano, riding a bicycle, or recognizing a face, is not a matter of following explicit rules but of having a bodily, perceptual attunement to the world. The classical AI assumption that intelligence can be captured in explicit rules running on a disembodied symbol manipulator is, from a Merleau-Pontyan perspective, a fundamental category error. Dreyfus's critique was initially dismissed by the AI community, but subsequent decades have increasingly vindicated many of his core insights. The difficulty of encoding common sense knowledge, the brittleness of rule based AI systems outside their training domains, and the successes of embodied and situated AI approaches all reflect the validity of the phenomenological challenge to AI's founding metaphysical assumptions. As Dreyfus (2007) observed in retrospect, the deep learning revolution, with its emphasis on learning from massive amounts of data rather than explicit rule following, can be seen as a partial but incomplete move away from the classical metaphysics, still limited by the absence of genuine embodiment and situatedness.

The Chinese Room and its Intentionality

John Searle's Chinese Room argument, presented in his landmark paper *"Minds, Brains, and Programs"* (1980), constitutes another fundamental challenge to the metaphysical assumptions of computationalist AI. In this case, to the assumption that syntax is sufficient for semantics, or that formal symbol manipulation is sufficient for genuine understanding. Searle's thought experiment asks us to imagine a person who does not understand Chinese locked in a room with a set of rules (a program) for manipulating Chinese symbols. The

person receives Chinese symbols as input and, by following the rules, produces Chinese symbols as output. To external observers, the person in the room appears to understand Chinese. But, Searle argues, the person understands nothing; they are merely manipulating symbols according to syntactic rules, with no comprehension of what the symbols mean. The argument is intended to show that a computer running a program, however sophisticated, is in the same position: it manipulates symbols according to syntactic rules, but this syntactic manipulation does not give rise to genuine semantic content or intentionality.

The metaphysical assumption that Searle challenges is what he calls "strong AI": the view that appropriately programmed computers have cognitive states and that the program itself is constitutive of the mental (Searle, 1980). Searle counters with the claim that intentionality, the property of being "about" something, is a biological phenomenon that cannot be reduced to formal computation. His theory of "biological naturalism" holds that mental states are caused by specific neurobiological processes and that no amount of symbol manipulation can duplicate this causal power. The responses to Searle's argument from defenders of AI, such as Daniel Dennett, Paul Churchland, Patricia Churchland, and Douglas Hofstadter, reflect the depth of the metaphysical disagreement at stake. The "systems reply" to Searle argues that while the person in the room does not understand Chinese, the system as a whole does. The "brain simulator reply" suggests that if the program simulated the actual causal structure of the brain, it would produce genuine understanding. These responses reveal that the debate about the Chinese Room is fundamentally a debate about the metaphysics of mind, about what is necessary and sufficient for intentionality, consciousness, and genuine understanding. These questions cannot be resolved by empirical AI research alone; they require extended metaphysical analysis.

Embodiment, Situatedness and New Metaphysical Frameworks

The limitations of classical AI's metaphysical assumptions have motivated the development of alternative approaches that draw on different philosophical frameworks. The embodied cognition movement, associated with researchers such as Rodney Brooks, Rolf Pfeifer, and Andy Clark, challenges the disembodied, representation-centric metaphysics of classical AI and draws instead on phenomenology, dynamical systems theory, and ecological psychology (Brooks, 1991; Pfeifer & Scheier, 1999; Clark, 1997). Rodney Brooks, in his influential paper *"Intelligence without Representation"* (1991), argued that intelligence arises not from internal representations of the world but from the direct coupling of a physical agent with its environment. His subsumption architecture dispensed with the central knowledge base of classical AI in favor of a layered system of behavioral modules directly responsive to environmental inputs. The metaphysical shift here is significant: from a picture of the mind as an inner theater of representations to a picture of intelligence as an emergent property of agent-environment interaction.

Andy Clark's *Being There: Putting Brain, Body, and World Together Again* (1997) and *Supersizing the Mind* (2008) develop what Clark calls the "extended mind" thesis, the view that the boundaries of the cognitive system extend beyond the skin of the organism to encompass environmental resources (Clark, 1997, 2008). This marks a radical departure from Cartesian metaphysics, which locates mind firmly within the skull. For Clark, the mind is not contained within the body but is distributed across the body-environment system. This metaphysical reorientation has implications for AI: it suggests that genuinely intelligent systems may need to be embedded in rich physical and social environments, not merely simulated in isolation. Francisco Varela, Evan Thompson, and Eleanor Rosch, in *The Embodied Mind: Cognitive Science and Human Experience* (1991), developed the concept of "enaction," the view that cognition is not the representation of a pre-given world but the bringing forth of a world through the embodied action of a living organism. This framework, drawing on Maurice Merleau-Ponty and Buddhist philosophy, represents perhaps the most thoroughgoing alternative to the metaphysics of classical AI. Intelligence, on the enactivist view, is inseparable from life: it is constituted by the self-organizing, sense-making activity of living systems in their environments. The implications for AI are profound: on this view, genuine artificial intelligence would require not merely simulation but artificial life, something that current AI systems lack.

Contemporary Deep Learning and Metaphysical Continuities

The rise of deep learning, particularly the development of deep neural networks, large language models, and generative AI systems, might seem to represent a clean break from the metaphysical assumptions of classical AI. Deep learning systems do not rely on hand-coded symbolic representations; they learn statistical patterns from massive datasets through gradient descent. They do not embody explicit rules or logical inference; they approximate complex functions through layers of non-linear transformations. In this sense, they appear to escape the formalizability thesis and the propositional knowledge assumption that Dreyfus

and others criticized (Dreyfus, 1992, 2007). Yet deep learning embeds its own metaphysical assumptions, no less contestable than those of classical AI. First, there is the assumption that intelligence can be reduced to pattern recognition over large datasets, that what we call understanding, reasoning, and knowledge is ultimately a form of sophisticated statistical correlation. This is a strongly empiricist and reductionist metaphysical assumption that learning theorists such as Vladimir Vapnik have made explicit in *The Nature of Statistical Learning Theory* (1995). Whether it is correct is not merely an empirical question but a metaphysical one: it depends on whether intelligence is, at bottom, pattern recognition or whether it involves something irreducibly more, something like genuine understanding, conceptual grasp, or intentionality.

Second, there is the assumption of scalability: that sufficiently large models trained on sufficiently large datasets will approximate and perhaps eventually achieve general intelligence. This assumption underlies the "scaling hypothesis" that has driven much recent AI research, as articulated by researchers at OpenAI and elsewhere. The metaphysical presupposition is that general intelligence is continuous with, and not categorically different from, the kind of pattern matching that neural networks perform. This is a substantive and contested metaphysical claim, related to the philosophical question of whether there are qualitative differences between different levels of cognitive organization or whether cognition is quantitatively continuous. Third, the black-box character of deep neural networks raises metaphysical questions about the nature of the representations they develop. Unlike classical AI, where representations are explicitly defined by the programmer, neural networks develop internal representations that are opaque even to their designers. What is the ontological status of these representations? Are they genuine semantic representations, mere statistical artifacts, or something else? Philosophers such as Gary Marcus and Yoshua Bengio have engaged with these questions from different perspectives. Marcus, in *Rebooting AI* (2019, co-authored with Ernest Davis), argues that deep learning lacks genuine understanding, while Bengio emphasizes the importance of causal and structured representations for achieving robust artificial intelligence. These debates are, at their root, metaphysical debates about the nature of representation and understanding. This, however give room to the doyen of metaphysical foundations of artificial intelligence

Ethics, Agency, and the Metaphysics of Artificial Agents

The question of whether AI systems can be genuine agents, with beliefs, intentions, desires, and perhaps moral status, is irreducibly metaphysical. It depends on prior answers to questions about the nature of agency, intentionality, consciousness, and moral considerability. The expanding capabilities of AI systems have made these metaphysical questions increasingly urgent (Dennett, 1987; Floridi, 2013). The debate about AI agency divides broadly into deflationary and inflationary positions. Deflationary positions, associated with philosophers such as Daniel Dennett and Luciano Floridi, argue that AI systems can be agents in a functional or intentional stance sense without having "real" mental states. Floridi's information ethics and his concept of artificial agents in works such as *The Ethics of Artificial Intelligence* (2011) treat agency as a matter of degree, applicable to thermostats, software programs, and human beings along a continuum (Floridi & Sanders, 2004). The metaphysical assumption here is a thoroughgoing functionalism: agency is constituted by the right kind of functional organization, not by any intrinsic property of the system.

Inflationary positions, associated with philosophers who regard consciousness and intentionality as intrinsically non-functional properties, deny that current AI systems are genuine agents or moral patients in any morally significant sense. Thomas Nagel's landmark essay *"What Is It Like to Be a Bat?"* (1974) argues that consciousness essentially involves a subjective, first-person point of view that cannot be captured in objective, third-person functional descriptions (Nagel, 1974). If Nagel is right, then no functional description of an AI system can capture whether it is conscious, and the question of AI moral status becomes deeply uncertain. The metaphysical stakes of AI ethics are therefore exceptionally high. Whether AI systems can be held responsible for their actions, whether they deserve protection from harm, and whether they can possess genuine interests all depend on prior metaphysical commitments concerning the nature of agency, consciousness, and moral status. As Nick Bostrom argues in *Superintelligence: Paths, Dangers, Strategies* (2014), the development of sufficiently advanced AI may create entities whose metaphysical status is genuinely uncertain, and the ethical consequences of misunderstanding that status could be catastrophic (Bostrom, 2014). Artificial Intelligence therefore emerges not as the negation of metaphysics but as one of its greatest contemporary expressions. Every AI system embodies assumptions concerning the nature of intelligence, reality, causation, representation, identity, and agency. Consequently, AI in this paper is it best understood as applied metaphysics translated into computational architecture.

On the Metaphysical Foundation of Artificial Intelligence

Artificial Intelligence (AI) is frequently portrayed as the product of advances in computer science, mathematics, engineering, and data science. While these disciplines provides the technical machinery for AI, they do not constitute its deepest intellectual foundation. Beneath every algorithm, neural network, and intelligent system lies a set of metaphysical assumptions concerning the nature of reality, mind, intelligence, causation, and existence. Indeed, AI would have been inconceivable without centuries of metaphysical inquiry into what thinking is, whether intelligence can exist apart from biological organisms, and whether reasoning can be formalized into universal principles. In this sense, metaphysics is not an external philosophical commentary on AI but the very foundation upon which the discipline rests. Historically, the roots of AI extend to classical Greek philosophy. Plato maintained that reality possesses an intelligible order accessible through reason rather than mere sensory experience. His Theory of Forms implied that universal truths exist independently of individual minds and that rational inquiry enables human beings to apprehend these eternal realities (Plato, 1992). This conviction that reality is rationally structured laid the earliest philosophical groundwork for the possibility that intelligence itself could be systematically understood. Likewise, Aristotle developed formal logic through the syllogism, demonstrating that valid reasoning follows objective principles independent of individual opinion (Aristotle, 1928). Aristotelian logic later became the intellectual ancestor of syllogistic logic, computation, and algorithmic reasoning, all of which are logical underpinnings of modern AI. From the metaphysical assumption of Aristotle comes the claim that rationality is ordered and therefore knowable. The world, according to him, has intelligible structures that the mind can grasp and replicate formally. This empiricist leaning nevertheless provided the software or blueprint for machines to reason. More so, his syllogistic logic influenced symbolic AI. It models intelligence as something derivable from fixed rules, a strong assumption in classical Ai knowledge.

Thomas Hobbes's description of reason as computation, the addition and subtraction of ideas or consequences, is often seen as phantasmal mental images originating from sensory perception. Hobbes, however, in my view, regards humans mechanistically, as automata governed by material processes. This, of course, supports viewing thought as algorithmic operations on representations, which is foundational to symbolic AI and, in extension, other cognitive science. Hobbes's metaphysical view is that the mind and thought are fundamentally material and mechanical; hence, no immaterial soul is necessarily required for reasoning. This, however, paves the way for the belief that a machine could replicate or exceed human cognition if it performs equivalent computations. His logic is that, if thinking is reckoning, then a sufficiently complex machine could also think.

The metaphysical transition from classical philosophy to modern AI became clearer during the seventeenth century through the rationalist tradition. Rene Descartes argued that the essence of human existence is thinking (*cogito, ergo sum*), while distinguishing between mind (*res cogitans*) and matter (*res extensa*) (Descartes, 1641/1998). Although Descartes himself doubted that machines could genuinely think, his separation of thought from biological substance unintentionally created the conceptual possibility that thinking might exist independently of the human body. Once thinking was conceived as an activity distinguishable from its material substrate, philosophers and scientists could begin asking whether non-biological systems might also instantiate intelligence. This intellectual trajectory was further strengthened by Gottfried Wilhelm Leibniz, whose dream of a *characteristica universalis* (universal symbolic language) and *calculus ratiocinator* (logical calculus) represented one of the earliest visions of automated reasoning (Leibniz, 1989). Leibniz believed that all rational disputes could eventually be resolved through symbolic calculation. His proposal anticipated modern symbolic AI by several centuries and demonstrated the metaphysical conviction that reason itself possesses a logical and computational structure. Contemporary computer programming, formal logic, and automated theorem proving all reflect this Leibnizian aspiration. His advances in binary notation and simple, perception-driven substances (Monads) prefigured digital computers. His metaphysical assumption that reality is rationally ordered and harmonious, pre-established by God, suggests that knowledge is compositional and calculable. Monads imply distributed, perceptual intelligence, and this underpins optimism towards coding and computing intelligence. The emergence of mathematical logic during the nineteenth and twentieth centuries further transformed metaphysical speculation into computational possibility. The symbolic logic developed by George Boole reduced logical reasoning to algebraic operations, thereby establishing that thought could be expressed through formal symbols (Boole, 1854). Later, the work of Alan Turing demonstrated that any process governed by explicit logical rules could, in principle, be executed by a machine. Turing's concept of the Universal Machine was therefore not merely an engineering innovation but also the culmination of a long metaphysical tradition that regarded reasoning as rule-governed and formalisable.

Conclusion

The foregoing analysis argues that artificial intelligence is not a metaphysically neutral enterprise. From its Cartesian roots in the formalizability of thought, through the functionalist and computationalist frameworks that gave it philosophical grounding, to the physicalist ontology that guides its contemporary development and the epistemological assumptions about knowledge and representation that shape its methods, AI is saturated with metaphysical commitments. These commitments determine what questions AI researchers ask, what solutions they consider possible, what counts as success or failure, and what ethical issues arise.

The phenomenological critiques of Dreyfus and Merleau-Ponty, the intentionality-based challenge of Searle, and the embodied and enactivist alternatives of Brooks, Clark, and Varela et al. all reveal the contestability of AI's founding assumptions and point toward alternative metaphysical frameworks that may open new possibilities for the field. Contemporary deep learning, while departing from the explicit formalism of classical AI, embeds its own metaphysical assumptions about the reducibility of intelligence to pattern recognition, the scalability of learning, and the nature of neural representations that are no less in need of philosophical scrutiny, which I think is the mistaken metaphysics of mind on their parts. As AI systems become increasingly powerful and pervasive, the stakes of getting the metaphysics right or at least of making the metaphysical assumptions explicit and subject to rational examination become correspondingly higher. The philosophy of AI is not a luxury or an afterthought; it is a necessary condition for a genuinely critical and self-aware AI research community. Indeed, AI rests on the age-old philosophical tradition that strongly demystified reason as something formalizable and mechanical, thereby making artificial replication something conceivable.

ORCID iD

Franklin Chibuzor Alo https://orcid.org/0009-0007-7085-4836

References

Aristotle. (1928). *Organon* (W. D. Ross, Trans.). Clarendon Press.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge.

Boole, G. (1854). *An investigation of the laws of thought.* Walton and Maberly.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain.* MIT Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* MIT Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* Oxford University Press.

Dennett, D. C. (1987). *The intentional stance.* MIT Press.

Dennett, D. C. (1991). *Consciousness explained.* Little, Brown and Company.

Descartes, R. (1998a). *Discourse on the method of rightly conducting one's reason and of seeking truth in the sciences* (D. A. Cress, Trans.). Hackett. (Original work published 1637)

Descartes, R. (1998b). *Meditations on first philosophy* (D. A. Cress, Trans.). Hackett. (Original work published 1641)

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason.* Harper & Row.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason.* MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence, 171*(18), 1137–1160.

Floridi, L. (2013). *The ethics of information.* Oxford University Press.

Floridi, L. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Fodor, J. A. (1975). *The language of thought.* Thomas Y. Crowell.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* MIT Press.

Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Leibniz, G. W. (1989). *Philosophical essays* (R. Ariew & D. Garber, Trans.). Hackett.

Lenat, D. B., & Guha, R. V. (1990). *Building large knowledge-based systems.* Addison-Wesley.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust.* Pantheon Books.

Merleau-Ponty, M. (1962). *Phenomenology of perception* (C. Smith, Trans.). Routledge. (Original work published 1945)

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence.* MIT Press.

Plato. (1992). *The Republic* (G. M. A. Grube, Trans.; C. D. C. Reeve, Rev.). Hackett.

Polanyi, M. (1966). *The tacit dimension.* Doubleday.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind.* New York University Press.

Putnam, H. (1975). The nature of mental states. In *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 429–440). Cambridge University Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* MIT Press.

von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior* (3rd ed.). Princeton University Press.

Whitehead, A. N. (1978). *Process and reality: An essay in cosmology.* Free Press. (Original work published 1929)