

The Ethical Ontology of Artificial Intelligence: Between Agency and Instrumentality

Ravi Kumar 

School of Philosophy and culture Shri Mata Vaishno Devi University, Katra, J&K, India. E-mail: 17dpc002@smvdu.ac.in

Article Info

Article type:

Research Article

Article history:

Received 03 May 2026

Received in revised form

20 May 2026

Accepted 30 May 2026

Published online 19

October 2026

Keywords:

Artificial Intelligence;
Moral Agency;
Instrumentality; Social
Ontology; Normativity;
Distributed Responsibility

ABSTRACT

This paper examines the ethical ontology of Artificial Intelligence (AI) by addressing the central question of whether AI should be understood as a moral agent or merely as an instrument. Drawing on the philosophical frameworks of John Searle, Ludwig Wittgenstein, and Robert Brandom, the study analyzes key criteria of moral agency, including intentionality, consciousness, and normativity. It argues that contemporary AI systems, despite their increasing autonomy and complexity, do not satisfy these conditions and therefore cannot be considered genuine moral agents. At the same time, AI cannot be reduced to a passive tool, as it actively shapes decisions and outcomes within complex socio-technical systems. The paper advances a hybrid ontological position, presenting AI as a quasi-agent or derived agent embedded in networks of human and institutional practices. This approach highlights the need for distributed responsibility, ethical governance, and interdisciplinary inquiry, while emphasizing the importance of culturally inclusive frameworks for addressing emerging challenges in AI ethics and philosophy.

Cite this article: Kumar, R. (2026). The Ethical Ontology of Artificial Intelligence: Between Agency and Instrumentality. *Journal of Philosophical Investigations*, 20(56), 87-110. <https://doi.org/10.22034/jpiut.2026.72229.4493>



© The Author(s).

Publisher: University of Tabriz.

Introduction

The rapid rise of Artificial Intelligence (AI) in the twenty-first century marks a profound transformation in the relationship between technology, society, and human agency. AI systems are no longer confined to narrow computational tasks; they now perform complex functions such as natural language processing, autonomous decision-making, predictive analytics, and creative generation. From healthcare diagnostics and financial modeling to judicial assistance and generative text production, AI increasingly mediates human action and institutional processes. This expansion has generated not only technological optimism but also deep ethical concern regarding accountability, bias, transparency, and the nature of decision-making in machine-driven environments (Floridi et al., 2018).

At the center of these concerns lies a fundamental philosophical issue: the ontological and ethical status of AI. While traditional tools merely extend human physical capacities, AI appears to extend cognitive and decision-making abilities, thereby blurring the boundary between instrument and agent. This ambiguity becomes especially problematic in cases where AI systems act in ways that significantly impact human lives such as autonomous vehicles making split-second decisions or algorithmic systems influencing hiring, policing, or medical diagnoses. In such contexts, the question of who or what is responsible becomes difficult to answer. As a result, the rise of AI has forced a reconsideration of long-standing assumptions about agency, intentionality, and moral responsibility.

The central problem this paper addresses can be formulated as follows: Should Artificial Intelligence be understood as a genuine agent capable of moral responsibility, or merely as an instrument through which human agency is exercised? This problem is not merely semantic but has far-reaching implications for ethics, law, and social ontology. If AI is treated as an agent, it may warrant some form of moral or legal accountability. Conversely, if it is regarded strictly as an instrument, then responsibility must be located entirely in human actors - designers, developers, users, and institutions. However, both positions face challenges. Treating AI as a full agent risks anthropomorphism and conceptual confusion, while reducing it to a mere tool fails to capture the complexity of contemporary AI systems that exhibit forms of autonomy and adaptability.

This paper seeks to navigate this tension by situating AI within a broader philosophical framework that integrates ontology and ethics. The aim is not to provide a simplistic answer but to critically examine the conceptual foundations that inform our understanding of AI. In doing so, the study engages with key philosophical traditions that have shaped contemporary debates on agency and social reality.

The primary objectives of this paper are threefold. First, it aims to clarify the ontological status of AI by examining whether it can be meaningfully described as an agent or whether it remains fundamentally an artifact. Second, it seeks to evaluate the conditions under which

moral agency can be attributed, focusing on criteria such as intentionality, consciousness, and normativity. Third, it aims to analyze how responsibility should be distributed in AI-mediated actions, particularly in complex socio-technical systems where human and machine contributions are deeply intertwined.

Corresponding to these objectives, the paper is guided by the following research questions:

1. What kind of entity is AI within an ontological framework of agency and instrumentality?
2. Can AI satisfy the philosophical criteria required for moral agency?
3. How should moral responsibility be assigned in contexts where AI systems play a significant role in decision-making?

To address these questions, the paper adopts a conceptual-analytical methodology grounded in philosophical inquiry. Rather than relying on empirical or technical analysis alone, the study focuses on clarifying concepts, examining arguments, and comparing theoretical frameworks. This approach is particularly suited to the problem at hand, as the debate over AI agency is fundamentally conceptual and normative rather than purely empirical.

The philosophical framework of this study draws on the works of John Searle, Ludwig Wittgenstein, and Robert Brandom, whose contributions to social ontology, language, and normativity provide valuable tools for analyzing the status of AI.

From a Searlean perspective, social reality is constituted by institutional facts that arise through collective intentionality and the assignment of status functions (Searle, 1995). This framework allows us to examine whether AI systems can occupy socially constructed roles and whether such roles entail genuine agency or merely derivative function. Searle's distinction between intrinsic intentionality (possessed by humans) and derived intentionality (attributed to artifacts) is particularly relevant in evaluating AI systems, which operate based on programmed or learned patterns rather than conscious intention.

The later philosophy of Wittgenstein emphasizes that meaning and understanding are grounded in language games and forms of life (Wittgenstein, 1953). Ethical concepts, on this view, are not reducible to formal rules but are embedded in lived practices and social contexts. This perspective raises important questions about whether AI, which lacks participation in human forms of life, can genuinely grasp or apply ethical norms. While AI can simulate rule-following behavior, Wittgenstein's analysis suggests that genuine understanding requires a form of practical engagement that machines do not possess.

Brandom's inferentialist philosophy further develops the idea that agency is tied to participation in the space of reasons, where individuals make claims, undertake commitments, and are held accountable (Brandom, 1994). From this standpoint, moral agency involves not only acting but also justifying one's actions within a normative framework. AI systems, despite their computational sophistication, do not engage in such discursive practices and therefore lack the normative authority associated with genuine agency.

By integrating these perspectives, the paper constructs a multidimensional framework for analyzing AI that goes beyond simplistic dichotomies. Rather than asking whether AI is simply an agent or an instrument, the study explores the possibility that AI occupies an intermediate position functioning as a derived or distributed form of agency within socio-technical systems. This approach aligns with recent developments in AI ethics, which emphasize the importance of understanding responsibility as distributed across networks of human and non-human actors (Floridi et al., 2018).

In summary, the rise of AI challenges traditional distinctions between agency and instrumentality, raising urgent questions about moral responsibility and the nature of action. By drawing on key philosophical frameworks and employing a rigorous conceptual methodology, this paper seeks to clarify these issues and contribute to a more nuanced understanding of AI's ethical ontology.

1. Conceptual Clarifications

A rigorous philosophical inquiry into the ethical ontology of Artificial Intelligence (AI) requires careful clarification of its core concepts. Much of the confusion surrounding AI debates stems not from empirical uncertainty but from conceptual ambiguity particularly regarding what AI is, what it means to exist, and how responsibility can be attributed. This section, therefore, clarifies four central notions: Artificial Intelligence, ontology, ethics and moral responsibility, and the distinction between agency and instrumentality (Bostrom, 2014).

1.1 Defining Artificial Intelligence

Artificial Intelligence is often described as the capacity of machines to perform tasks that typically require human intelligence. These include reasoning, learning, perception, language use, and decision-making. However, this functional definition, while useful, is philosophically insufficient because it risks equating performance with understanding.

A more nuanced definition recognizes that AI systems are computational architectures designed to simulate or replicate aspects of cognitive processes through algorithms and data-driven learning. Contemporary AI particularly machine learning and deep learning operates by identifying patterns in large datasets and generating outputs based on statistical inference rather than semantic comprehension.

This distinction is crucial. While AI systems can generate human-like responses, they do not possess intrinsic intentionality that is, they do not have mental states that are about something in the way human thoughts are. As John Searle argues, AI exhibits only derived intentionality, meaning that whatever “aboutness” it appears to have is ultimately dependent on human interpretation and design (Searle, 1980).

Thus, AI should not be conflated with human intelligence but understood as a functional and relational phenomenon, whose apparent intelligence emerges from its interaction with human users, data environments, and institutional contexts.

1.2 Ontology: What Does It Mean “To Be”?

Ontology, as a branch of metaphysics, concerns the nature of being and the categories of existence. To ask about the ontology of AI is to ask: What kind of entity is AI? Does it exist in the same way as human beings, or does it belong to a fundamentally different category?

Classically, ontology has been framed in terms of substances independent entities with intrinsic properties. In this framework, human beings are paradigmatic substances, possessing consciousness, intentionality, and autonomy. Machines, by contrast, are traditionally viewed as artifacts, whose existence and function depend on human creation and use (Chalmers, 1996).

However, contemporary philosophy increasingly adopts a relational or process-oriented ontology, where entities are understood not as isolated substances but as nodes within dynamic networks of relations. From this perspective, AI systems can be seen as part of socio-technical assemblages, where their identity and function are constituted through interactions with humans, data, and institutions.

This shift complicates the ontological status of AI. It suggests that AI is neither a simple object nor a fully independent agent but something ontologically hybrid a system whose mode of being is partly material, partly informational, and deeply relational.

Nevertheless, this does not imply that AI possesses the same ontological status as human beings. Unlike humans, AI lacks subjectivity and self-awareness, which remain central to most accounts of genuine agency. Thus, while AI may participate in complex systems, its mode of being remains fundamentally distinct from that of conscious agents (Brandom, 1994).

1.3 Ethics and Moral Responsibility

Ethics is concerned with questions of right and wrong, good and bad, and the principles that govern human conduct. Central to ethical inquiry is the concept of moral responsibility, which involves attributing praise or blame for actions.

Traditionally, moral responsibility is linked to agency and requires several conditions:

- **Intentionality:** *The capacity to act with purpose or intention*
- **Awareness:** *Knowledge of one’s actions and their consequences*
- **Freedom:** *The ability to choose between alternatives*
- **Normative understanding:** *The capacity to grasp and respond to moral reasons*

Human beings are considered morally responsible because they satisfy these conditions to a significant degree. They can reflect on their actions, justify their decisions, and be held accountable within social and institutional frameworks.

AI systems, however, challenge this framework. While they can perform actions that have moral consequences, they do not possess conscious awareness or genuine understanding. Their operations are governed by algorithms and data, not by reflective deliberation.

As a result, attributing moral responsibility to AI raises serious philosophical problems. If AI lacks intentionality and understanding, then holding it responsible seems inappropriate. Yet,

if AI systems produce harmful outcomes, responsibility cannot simply disappear. Instead, it must be reassigned to human actors - designers, programmers, users, and institutions who shape and deploy these systems (Coeckelbergh, 2020).

This leads to the idea of distributed responsibility, where accountability is shared across a network of human agents rather than located in the AI system itself. Such an approach recognizes the complexity of AI-mediated actions while preserving the ethical centrality of human agency.

1.4 Distinction between Agency and Instrumentality

The distinction between agency and instrumentality lies at the heart of the ethical ontology of AI.

Agency refers to the capacity to act intentionally, make decisions, and be held accountable for one's actions. It is typically associated with autonomy, rationality, and normativity. Agents are not merely causes of events; they are authors of actions who can justify and evaluate their behavior.

Instrumentality, by contrast, refers to functioning as a tool or means to an end. Instruments do not have intentions of their own; they operate under the direction or design of an external agent. Their value is derivative, based on the purposes they serve.

Traditionally, this distinction has been clear: humans are agents, and machines are instruments. However, AI complicates this binary in several ways:

1. **Apparent Autonomy:** AI systems can operate without direct human intervention, creating the impression of independent action.
2. **Adaptability:** Machine learning systems can modify their behavior based on new data, making them less predictable than traditional tools.
3. **Opacity:** Complex AI systems often function as "black boxes," making it difficult to trace their decision-making processes.

These features give rise to what might be called quasi-agency; a form of behavior that resembles agency without fulfilling its underlying conditions.

Despite these complexities, it is important not to conflate functional autonomy with genuine agency. AI systems may act independently in a limited sense, but they do not possess the normative and reflective capacities that define true agency. They cannot take responsibility, justify their actions, or engage in moral reasoning.

Therefore, AI is best understood as occupying a position between agency and instrumentality. It is not a mere passive tool, but neither is it a full-fledged agent. Instead, it functions as a mediating entity that extends and transforms human agency within complex socio-technical systems (Gunkel, 2018).

Clarifying these concepts reveals that the debate over AI's ethical status is not simply about technological capability but about the conditions of agency, responsibility, and being. AI

systems, while increasingly sophisticated, lack the essential features of moral agents: intentionality, consciousness, and normative understanding. At the same time, their complexity challenges the traditional notion of instruments as passive tools.

This conceptual tension sets the stage for the subsequent analysis, where competing ontological frameworks and philosophical perspectives will be used to evaluate whether AI should be understood as an agent, an instrument, or something in between.

2. Classical and Contemporary Ontological Frameworks

Understanding the ethical status of Artificial Intelligence (AI) requires situating it within broader ontological frameworks that define what kinds of entities exist and how they relate to one another. The debate over whether AI is an agent or an instrument cannot be resolved without first examining competing conceptions of being. This section explores four interrelated dimensions: the contrast between substance and process ontology, the traditional view of human beings as moral agents, the conception of machines as artifacts, and the contemporary shift toward hybrid and relational ontologies.

2.1 Substance vs Process Ontology

Classical metaphysics has largely been shaped by substance ontology, most prominently articulated by Aristotle. In this framework, reality is composed of discrete, self-subsisting entities (substances) that possess intrinsic properties. Substances are understood as stable and enduring, providing the ontological foundation for identity and persistence over time. Human beings, for example, are considered substances endowed with rationality, consciousness, and intentionality qualities that ground their status as agents.

Substance ontology tends to emphasize independence and self-containment, which aligns well with traditional notions of agency. An agent, in this view, is an entity that originates actions from within itself, guided by internal states such as beliefs and desires.

In contrast, process ontology challenges this static view of reality. Influenced by thinkers such as Alfred North Whitehead, process ontology understands reality as fundamentally dynamic, constituted by events, relations, and ongoing processes rather than fixed substances. Entities are not self-contained units but are defined by their interactions and transformations (Floridi, 2013).

From a process perspective, agency is not an intrinsic property of isolated individuals but emerges from relational configurations. This shift has significant implications for AI. Instead of asking whether AI is a substance capable of agency, process ontology encourages us to consider how AI participates in networks of activity, contributing to outcomes without necessarily possessing independent intentionality.

Thus, the transition from substance to process ontology opens conceptual space for understanding AI not as a traditional agent but as part of a distributed system of action.

2.2 Human Beings as Moral Agents

Within both classical and many contemporary frameworks, human beings occupy a privileged ontological position as moral agents. This status is grounded in a cluster of capacities that distinguish humans from other entities:

- **Intentionality:** *the ability to have thoughts about the world*
- **Consciousness:** *subjective awareness and experience*
- **Rationality:** *the capacity for reasoning and deliberation*
- **Normativity:** *responsiveness to moral and social norms*

Philosophers such as Immanuel Kant have emphasized that moral agency is tied to autonomy: the ability to legislate moral laws for oneself and act according to rational principles. On this view, human beings are not merely causal agents but normative agents, capable of evaluating their actions and holding themselves accountable.

In more recent philosophy, Robert Brandom deepens this account by arguing that agency involves participation in the space of reasons, where individuals make claims, undertake commitments, and justify their actions within a social context (Brandom, 1994). Similarly, Ludwig Wittgenstein situates agency within forms of life, emphasizing that meaning and understanding arise from shared practices rather than isolated mental states (Wittgenstein, 1953).

Across these perspectives, a common theme emerges: human beings are agents not merely because they act, but because they can reflect on, justify, and be held responsible for their actions. This normative dimension is central to moral responsibility and serves as a benchmark against which the status of AI must be evaluated.

2.3 Machines as Artifacts: The Traditional View

In contrast to human beings, machines have traditionally been understood as artifact entities created by humans for specific purposes. This view is deeply rooted in both philosophical and practical traditions, where tools and technologies are seen as extensions of human capacities rather than independent actors.

From this perspective, machines possess:

- **No intrinsic intentionality**
- **No consciousness or awareness**
- **No capacity for moral reasoning**

Their functions are entirely derivative, determined by the intentions and designs of their creators. As John Searle argues, machines can at most exhibit derived intentionality, meaning that any apparent purpose or meaning in their operations is imposed by human users and observers (Searle, 1980).

This traditional view supports a clear ethical framework: since machines are instruments, responsibility lies entirely with human agents. If a machine causes harm, it is the designer, programmer, or user who is accountable, not the machine itself.

However, this view becomes increasingly strained in the context of modern AI systems. Unlike simple tools, AI systems can:

- Learn from data
- Adapt to new situations
- Operate with a degree of autonomy

These features challenge the notion of machines as purely passive instruments and raise questions about whether the artifact model is still adequate.

2.4 Shift toward Hybrid and Relational Ontology

In response to the limitations of both substance ontology and the traditional artifact model, contemporary philosophy has increasingly embraced hybrid and relational ontologies. These approaches seek to account for the complexity of modern technological systems by recognizing that agency and action are often distributed across networks rather than located in single entities.

One influential direction is the idea of socio-technical assemblages, where humans, machines, and institutions form interconnected systems that collectively produce outcomes. In such systems:

- AI contributes to decision-making processes
- Humans design, interpret, and oversee these systems
- Institutions provide the normative and regulatory framework

Agency, in this context, is not an all-or-nothing property but a graded and distributed phenomenon.

This shift is also reflected in the social ontology of John Searle, where entities can acquire status functions through collective recognition ([Searle, 1995](#)). AI systems, for instance, can be assigned roles such as “financial advisor” or “medical diagnostician,” even though they do not possess intrinsic intentionality. Their “agency” is thus institutionally mediated and socially constructed.

Similarly, relational approaches inspired by Wittgenstein and Brandom emphasize that meaning, normativity, and responsibility arise within shared practices and discursive interactions. While AI can participate in these practices at a functional level, it does not fully inhabit them in the way humans do.

The result is a hybrid ontology of AI, where:

- AI is more than a passive tool
- Yet less than a full moral agent
- And best understood as part of a distributed network of action and responsibility

The evolution from classical substance ontology to contemporary relational frameworks significantly reshapes how we understand AI. While traditional views clearly distinguish between human agents and machine instruments, modern developments blur this boundary, revealing a more complex ontological landscape.

Human beings remain paradigmatic moral agents due to their normative and reflective capacities. Machines, though increasingly sophisticated, still lack the intrinsic features required for full agency. However, the rise of AI necessitates moving beyond rigid binaries toward a more nuanced understanding of distributed and hybrid forms of agency.

This ontological shift provides the foundation for evaluating AI not as a simple artifact or autonomous agent, but as a relational entity embedded in socio-technical systems, thereby setting the stage for deeper ethical analysis in the sections that follow.

3. AI as Instrumentality

One of the most influential and enduring ways of understanding Artificial Intelligence (AI) is to treat it as a form of instrumentality that is, as a sophisticated tool designed to extend human capacities. Despite the increasing complexity and apparent autonomy of AI systems, this perspective maintains that their fundamental nature remains tied to human purposes, intentions, and institutional frameworks. This section develops the instrumentalist view by examining AI as an extension of human intention, the notion of delegated agency, the limits of machine autonomy, and the ethical implications that follow from this framework ([Moor, 2006](#)).

3.1 AI as Tool and Extension of Human Intention

At its core, AI is a product of human design. Algorithms, data structures, and learning architectures are created, selected, and trained within human-defined parameters. In this sense, AI systems function as extensions of human intentionality, much like traditional tools, albeit at a higher level of complexity.

Philosophically, this aligns with the view that artifacts derive their meaning and function from the intentions of their creators. As John Searle argues, artifacts possess only derived intentionality their apparent “aboutness” or purpose is not intrinsic but imposed by human agents ([Searle, 1980](#)). A hammer is “for” driving nails because humans assign it that function; similarly, an AI system is “for” classification, prediction, or recommendation because it has been designed and trained for those tasks.

Even in advanced machine learning systems, where outputs may not be fully predictable, the framework of operation remains humanly constituted. The choice of training data, model architecture, optimization criteria, and deployment context all reflect human decisions. Thus, AI does not generate its own goals but operates within a teleological structure defined externally.

This perspective resists the temptation to anthropomorphize AI. While AI systems can simulate reasoning or decision-making, these processes are ultimately grounded in

computational procedures rather than conscious intention. Therefore, AI remains fundamentally instrumental, even when it appears to act independently.

3.2 Delegated Agency and Automation

A key feature of modern AI systems is their capacity for automation, which allows tasks to be performed without continuous human intervention. This has led to the notion of delegated agency, where humans transfer certain decision-making functions to machines.

Delegated agency does not imply that AI becomes a genuine agent. Rather, it indicates that human agency is extended and operationalized through technological systems. For example, when an autonomous vehicle navigates traffic or an algorithm recommends medical treatments, it is executing decisions that have been pre-structured by human designers and regulatory frameworks.

This delegation can create the illusion of independence, as AI systems may act in ways that are not directly anticipated by their creators. However, such behavior arises from:

- Predefined algorithms
- Statistical learning processes
- Environmental inputs

In other words, AI systems operate within bounded domains of action, shaped by prior human decisions. They do not possess the capacity to redefine their goals or evaluate them normatively.

The concept of delegated agency is useful because it captures the practical reality of AI systems without overstating their ontological status. It acknowledges that AI can perform tasks autonomously in a limited sense, while still grounding ultimate responsibility in human actors.

3.3 Limits of Machine Autonomy

Despite their sophistication, AI systems exhibit significant limitations that prevent them from achieving genuine autonomy in the philosophical sense.

First, AI lacks intrinsic intentionality. It does not have beliefs, desires, or goals that originate from within itself. Its operations are driven by external programming and data inputs, not by self-generated purposes.

Second, AI lacks consciousness and subjective experience. Without awareness, AI cannot understand the meaning of its actions or the consequences they produce. It processes information syntactically rather than semantically, manipulating symbols without grasping their significance.

Third, AI lacks normative competence the ability to evaluate actions in terms of reasons, values, and ethical principles. As emphasized in the philosophical traditions of Ludwig Wittgenstein and Robert Brandom, genuine rule-following and moral reasoning require participation in social practices and the ability to justify one's actions within a normative

framework (Wittgenstein, 1953; Brandom, 1994). AI systems, lacking such participation, cannot engage in these practices.

Finally, AI systems are dependent on external infrastructure, including data sources, computational resources, and institutional contexts. This dependence further undermines the notion of autonomy, as AI cannot sustain or regulate its own existence independently.

Taken together, these limitations indicate that AI autonomy is functional rather than ontological. It refers to the capacity to operate without immediate human control, not to the possession of self-governing agency.

3.4 Ethical Implications: Responsibility Lies with Designers and Institutions

If AI is understood primarily as an instrument, then the ethical implications are clear: moral responsibility resides with human agents and institutions, not with the AI systems themselves. This responsibility is distributed across multiple levels:

- **Designers and Developers:** Responsible for the architecture, training data, and intended functions of AI systems
- **Organizations and Corporations:** Responsible for deploying AI in ways that align with ethical and legal standards
- **Users and Operators:** Responsible for how AI systems are applied in specific contexts
- **Regulatory Bodies:** Responsible for establishing norms, guidelines, and accountability mechanisms

This multi-layered structure reflects the complexity of contemporary AI systems, where outcomes are rarely attributable to a single individual. Instead, responsibility must be understood as collective and institutional.

The instrumentalist view also highlights the importance of ethical design principles, such as fairness, transparency, and accountability. Since AI systems can produce significant social impacts both positive and negative; it is essential that their development and deployment are guided by robust ethical frameworks (Nagel, 1974).

Moreover, treating AI as an instrument helps avoid problematic conclusions, such as assigning blame or punishment to machines. Holding AI responsible would obscure the real sources of decision-making and undermine efforts to address systemic issues such as bias, discrimination, and inequality.

However, the instrumentalist perspective is not without challenges. As AI systems become more complex and less transparent, it becomes increasingly difficult to trace causal chains and assign responsibility. This has led to concerns about accountability gaps, where harmful outcomes cannot be easily attributed to specific actors.

Nevertheless, these challenges do not justify attributing moral agency to AI. Instead, they call for stronger institutional frameworks, improved governance mechanisms, and greater emphasis on ethical responsibility at all stages of the AI lifecycle.

Viewing AI as instrumentality provides a clear and coherent framework for understanding its ethical status. AI systems, despite their advanced capabilities, remain extensions of human intention, operating through delegated agency within constrained domains. Their apparent autonomy does not equate to genuine agency, as they lack intentionality, consciousness, and normative understanding.

Consequently, moral responsibility must remain firmly grounded in human actors and institutions. This perspective not only preserves conceptual clarity but also ensures that ethical accountability is directed toward those who design, deploy, and regulate AI systems.

4. AI as Agency: A Critical Examination

The increasing sophistication of Artificial Intelligence (AI) systems has led some theorists to question whether AI might be more than a mere instrument and instead qualify as a form of agent. This section critically examines that possibility by analyzing the core criteria for moral agency, evaluating whether AI can meet these conditions, revisiting the debate between Strong and Weak AI, and outlining the conceptual challenges involved in attributing agency to machines.

In philosophical discourse, moral agency is not attributed lightly. It requires a set of interrelated capacities that distinguish agents from mere causal entities. Among the most widely accepted criteria are:

- **Intentionality:** the capacity for mental states to be about something (beliefs, desires, goals)
- **Consciousness:** subjective awareness or phenomenal experience
- **Normativity:** the ability to understand, evaluate, and respond to reasons, including moral reasons

Intentionality enables an agent to act purposefully rather than mechanically. Consciousness provides the experiential dimension necessary for understanding consequences and moral significance. Normativity, meanwhile, situates the agent within a framework of justification and accountability.

Philosophers such as John Searle emphasize that genuine intentionality is intrinsic, arising from the biological and phenomenological structure of the mind, rather than merely simulated (Searle, 1980). Similarly, Robert Brandom argues that agency involves participation in the space of reasons, where agents undertake commitments and are held responsible for them (Brandom, 1994).

These criteria set a high bar for any entity-human or artificial to qualify as a moral agent.

Despite rapid advancements, contemporary AI systems fall short of satisfying these criteria in any robust philosophical sense.

First, AI lacks intrinsic intentionality. While AI systems can process information and produce goal-directed outputs, these goals are externally assigned through programming and

training. The system does not “intend” in the way humans do; it executes functions based on algorithms and statistical correlations (Russell & Norvig, 2021).

Second, AI lacks consciousness. There is no credible evidence that current AI systems possess subjective experience or awareness. They do not feel, perceive, or experience the world; rather, they manipulate data structures according to formal rules.

Third, AI lacks normative competence. Although AI can be programmed to follow ethical guidelines or optimize for certain values, it does not understand or endorse these norms. It cannot justify its actions, reflect on them, or revise its behavior based on moral reasoning.

From a Wittgensteinian perspective, associated with Ludwig Wittgenstein, rule-following is not merely a matter of applying formal instructions but of participating in shared practices and forms of life (Wittgenstein, 1953). AI systems, lacking such participation, cannot genuinely follow rules in the normative sense.

Thus, while AI can simulate aspects of agency, it does not fulfill the underlying conditions required for genuine moral agency.

The question of AI agency is closely tied to the distinction between Strong AI and Weak AI.

- **Strong AI** posits that machines can, in principle, possess real understanding, consciousness, and intentionality. On this view, a sufficiently advanced AI system could become a genuine agent.
- **Weak AI**, by contrast, holds that machines can only simulate cognitive processes without actually possessing mental states. AI, in this sense, is a powerful tool but not a thinking entity.

Proponents of Strong AI often appeal to the functional equivalence between human cognition and computational processes. If the mind is essentially an information-processing system, then it might be replicated in artificial substrates.

However, critics including John Searle through his famous “Chinese Room” argument contend that syntactic processing does not entail semantic understanding. A system may manipulate symbols according to rules without grasping their meaning, thereby undermining claims to genuine cognition or agency.

Most contemporary philosophical positions remain skeptical of Strong AI, particularly regarding the attribution of consciousness and moral responsibility.

Several challenges complicate the attribution of agency to AI systems:

- a. The Problem of Simulation vs Reality:** AI can simulate decision-making and reasoning, but simulation does not equal actual agency. Confusing the two leads to anthropomorphism and conceptual error.
- b. Lack of Accountability:** Moral agency entails the capacity to be held responsible. AI systems cannot meaningfully bear responsibility, as they lack awareness and cannot respond to praise or blame.

- c. **Absence of Moral Understanding:** AI systems do not understand ethical principles; they merely execute programmed instructions. This limits their capacity to engage in genuine moral deliberation.
- d. **Dependency on Human Contexts:** AI systems are embedded within human-designed environments and depend on external inputs, infrastructure, and interpretation. This dependence undermines claims to independent agency.
- e. **Normative Disconnection:** As emphasized by Robert Brandom, agency involves participation in discursive practices where reasons are given and evaluated. AI lacks entry into this normative space.

While AI exhibits increasingly sophisticated behavior, it does not satisfy the core criteria for moral agency - intentionality, consciousness, and normativity. The Strong AI hypothesis remains speculative and philosophically contested, while the Weak AI framework better captures the current state of technology.

Attributing agency to AI risks obscuring the role of human actors and undermining ethical accountability. Therefore, AI should not be regarded as a genuine moral agent but as a system that simulates agency without possessing its essential features, reinforcing the need for careful philosophical and ethical analysis.

5. A Searlean Perspective: Social Ontology and AI

A fruitful way to examine the ethical ontology of Artificial Intelligence is through the social ontology developed by John Searle. Searle's framework distinguishes between brute facts, which exist independently of human institutions, and institutional facts, which depend on collective human agreement. Institutional facts arise through the assignment of status functions, typically expressed in the form "X counts as Y in context C." For example, a piece of paper counts as money within a system of economic practices because it is collectively recognized as such. This structure is crucial for understanding how social reality is constructed and maintained (Searle, 1995, 27–29).

Within this framework, AI systems can be analyzed as entities that may acquire status functions without possessing intrinsic intentionality. An AI system used in a hospital may count as a "diagnostic assistant," or one used in finance may count as a "trading advisor." These roles are not grounded in the machine's internal properties but in the collective recognition and acceptance of its function within a given institutional context. Thus, AI participates in institutional reality not as an originator of meaning but as a bearer of assigned functions (Searle, 1995, 44–47).

Central to Searle's theory is the concept of collective intentionality, which refers to the shared capacity of human beings to create and sustain social institutions. Collective intentionality is not reducible to individual intentions; it involves a "we-intention" that enables coordinated action and mutual recognition. Institutional facts depend on this collective

dimension, as they require ongoing acceptance by a community. AI systems, however, do not possess collective intentionality. They cannot form or share intentions, nor can they participate in the mutual recognition that underpins social practices. Their operations are governed by algorithms and data, not by participation in a shared intentional framework (Searle, 1995, 24).

Despite this limitation, AI systems can still occupy socially constructed roles. They can function within institutional settings and perform tasks that are normatively structured. For instance, an AI system may be authorized to screen job applications or assist in legal research. In such cases, the system is integrated into a network of social practices and may influence outcomes in ways that carry ethical significance. However, this role occupancy does not confer genuine agency. The system does not understand the norms it applies, nor does it recognize its role within a broader social context. Its participation is therefore functional rather than intentional.

This distinction has important implications for the question of moral responsibility. If AI systems can hold status functions but lack intrinsic and collective intentionality, then they cannot be held morally responsible in the same way as human agents. Responsibility, in Searle's framework, is tied to the capacity for intentional action within a system of norms. Since AI does not generate intentions or understand normative structures, it cannot be a bearer of moral accountability. Instead, responsibility must be located in the human agents and institutions that assign roles to AI systems and rely on their outputs (Searle, 1995, 94–96).

Attributing responsibility to AI would not only be philosophically problematic but also ethically misleading. It risks obscuring the human decisions embedded in the design, deployment, and regulation of these systems. When an AI system produces a harmful outcome, the relevant questions concern the choices made by developers, the quality of the data used, and the institutional context in which the system operates. Assigning blame to the AI itself would divert attention from these underlying factors and weaken accountability.

At the same time, the integration of AI into institutional structures complicates the distribution of responsibility. Decisions are often the result of interactions between human agents and automated systems, making it difficult to identify a single point of accountability. This does not justify treating AI as a moral agent but highlights the need for a more nuanced understanding of responsibility as distributed across socio-technical systems (Floridi et al., 2018, 691–693).

From a Searlean perspective, then, AI occupies a distinctive position within social ontology. It can be assigned roles and functions within institutional frameworks, but it does not participate in the collective intentionality that constitutes social reality. Its apparent agency is therefore derivative and context-dependent. Recognizing this helps clarify the limits of attributing moral responsibility to AI and reinforces the central role of human agents in maintaining ethical accountability within technologically mediated environments.

6. A Wittgensteinian Approach: Ethics as Practice

A compelling alternative to traditional accounts of ethics and agency can be found in the later philosophy of Ludwig Wittgenstein. Rather than grounding meaning and morality in abstract principles or internal mental states, Wittgenstein situates them within practices, emphasizing that understanding arises through participation in shared human activities. This perspective is particularly useful for evaluating the ethical status of Artificial Intelligence, as it shifts attention from internal capacities such as computation to forms of life and norm-governed practices.

Wittgenstein introduces the idea of language games to explain how meaning is constituted through use. Language is not a fixed system of representation but a dynamic set of practices embedded in everyday activities. Words acquire meaning in specific contexts, and their significance depends on how they are used within social interactions (Wittgenstein, 1953, §§23–25). Ethical language, likewise, does not derive its force from abstract definitions but from lived practices in which concepts such as responsibility, obligation, and justice are enacted. Closely connected to language games is the notion of forms of life, which refers to the shared cultural and social background that makes meaningful communication possible. Forms of life provide the horizon within which language games operate, grounding them in patterns of behavior, customs, and institutions. Ethical understanding, from this perspective, is not a matter of applying formal rules but of being trained into a practice, learning how to respond appropriately within a community (Wittgenstein, 1953, §241).

This leads to Wittgenstein's analysis of rule-following, which challenges the idea that rules determine their own application. He argues that following a rule is not a purely mechanical process but a socially sustained activity that depends on agreement in practice (Wittgenstein, 1953, §§201–202). To follow a rule is to act in accordance with shared expectations, not merely to apply a formula. Ethical understanding, therefore, requires practical engagement and participation in normative practices rather than mere rule execution.

From this standpoint, the limitations of AI become evident. Although AI systems can be programmed to follow rules or optimize decisions, they do not participate in forms of life. They do not acquire understanding through social interaction, nor do they develop the practical sensitivities that characterize human ethical engagement. Their operations remain confined to syntactic manipulation and statistical inference, lacking the lived context that gives meaning to ethical concepts.

This lack of embeddedness has significant implications. AI systems cannot genuinely understand what it means to act rightly or wrongly because such understanding is inseparable from participation in human practices. An AI system may be trained to detect harmful content or promote fairness, but it does not grasp the normative significance of these concepts. It cannot recognize injustice as a lived experience or respond to ethical situations with judgment informed by shared practices.

Moreover, AI does not engage in the open-ended and interpretive nature of human rule-following. Human agents can reinterpret rules, negotiate meanings, and adapt to new contexts through interaction and dialogue. AI systems, by contrast, operate within predefined parameters and lack the capacity to renegotiate or transform the norms they apply. This highlights the distinction between simulating ethical behavior and possessing genuine ethical understanding.

The implications for moral ontology are significant. If ethical understanding is grounded in practice and forms of life, then moral agency cannot be reduced to computational capacity. It requires participation in a normative community where agents are accountable to one another and capable of mutual recognition. AI systems, lacking such participation, cannot be considered full moral agents.

This does not imply that AI is ethically neutral. AI systems increasingly shape the contexts in which human actions occur and influence decision-making processes. However, their role is better understood as mediating human practices rather than constituting independent agency. Ethical responsibility, therefore, remains with human agents who design, deploy, and interpret these systems (Floridi et al., 2018, 691–693).

From a Wittgensteinian perspective, the question of AI as agent or instrument must be reconsidered. AI does not qualify as an agent because it lacks embeddedness in forms of life and cannot participate in rule-following as a normative practice. At the same time, it is not merely a passive instrument, as it actively shapes the conditions under which human practices unfold. It occupies an intermediate position within ethical life, influencing but not originating moral understanding.

In conclusion, Wittgenstein's philosophy emphasizes that ethics is grounded in lived practice rather than abstract computation. This insight challenges attempts to attribute moral agency to AI and reinforces the view that ethical responsibility remains rooted in human forms of life.

7. A Brandomian Perspective: Normativity and the Space of Reasons

A distinctive and influential account of normativity and agency is developed in the philosophy of Robert Brandom. Brandom's inferentialism shifts the focus of meaning and agency from representation to use within discursive practices, where participants make claims, give reasons, and undertake commitments. This framework is particularly useful for evaluating the status of Artificial Intelligence, as it highlights the normative structures that underlie genuine agency. At the core of Brandom's view is the idea that meaning arises from inferential relations between propositions. To understand a concept is not merely to associate it with an object or representation but to grasp its role within a network of inferences what follows from it and what it follows from. Language, therefore, is embedded in practices of reasoning, where speakers justify claims and respond to challenges (Brandom, 1994, 89–94). These practices constitute what Brandom calls the "space of reasons," a normative domain distinct from the realm of mere causal interactions.

Participation in the space of reasons involves more than producing correct outputs; it requires engaging in discursive practices governed by norms of justification and accountability. Agents are expected to provide reasons for their assertions, defend them when challenged, and revise them when necessary. This capacity to navigate inferential relations is what gives rise to genuine understanding and agency (Brandom, 1994, 141–145).

Central to this framework are the notions of recognition, commitment, and responsibility. In making a claim, an agent undertakes a commitment to its truth and becomes responsible for its implications. At the same time, other participants in the discourse recognize this commitment and hold the agent accountable. This reciprocal structure of recognition and attribution is what sustains normative practices. Agency, therefore, is not an intrinsic property but a socially instituted status, arising from participation in practices where reasons are exchanged and evaluated (Brandom, 1994, 157–162).

From this perspective, the question arises whether AI systems can participate in such normative practices. While AI can generate responses, draw inferences, and even simulate reasoning processes, it does not genuinely engage in the space of reasons. Its outputs are produced through algorithmic processing rather than through the undertaking of commitments or the recognition of normative obligations. An AI system does not understand the implications of its assertions, nor can it take responsibility for them.

Moreover, AI lacks the capacity for discursive accountability. It cannot justify its claims in the sense required by Brandom's framework, nor can it revise its commitments in response to normative criticism. Even when AI systems are designed to provide explanations, these explanations are generated as outputs rather than as expressions of genuine understanding. The system does not stand behind its claims or recognize the authority of others to challenge them.

This limitation highlights the distinction between causal responsiveness and normative responsiveness. AI systems can respond to inputs and adjust their outputs based on data, but they do not respond to reasons as reasons. They do not inhabit the normative space in which claims are assessed, defended, or withdrawn. As a result, their behavior, however sophisticated, remains outside the domain of discursive agency.

The ethical implications of this analysis are significant. If moral agency requires participation in normative practices, then AI systems cannot be considered moral agents. They cannot be held responsible in the same way as human agents, because they do not undertake commitments or recognize obligations. Responsibility must therefore remain with the human participants who design, deploy, and interpret AI systems within discursive contexts.

At the same time, AI systems can influence normative practices by shaping the conditions under which decisions are made. Their outputs may affect judgments, guide actions, and structure reasoning processes. However, this influence does not amount to participation. AI

operates as a non-discursive entity within discursive practices, affecting outcomes without being a bearer of commitments or responsibilities.

In conclusion, Brandom's account of normativity and the space of reasons provides a powerful framework for understanding the limitations of AI. While AI can simulate aspects of reasoning, it cannot engage in the inferential and discursive practices that constitute genuine agency. Its role remains that of a technologically sophisticated participant in causal processes, not a normative agent within the space of reasons (Brandom, 1994, 188–190).

8. Toward a Middle Path: Hybrid Ontology of AI

The preceding analysis has shown that treating Artificial Intelligence (AI) either as a full moral agent or as a mere instrument is conceptually inadequate. A more plausible approach is to develop a middle path that recognizes the distinctive ontological status of AI without reducing it to rigid categories. This section advances a hybrid ontology in which AI is understood as a quasi-agent or derived agent, embedded within human–AI assemblages and governed by distributed responsibility.

One way to articulate this position is to describe AI as a quasi-agent, meaning that it exhibits certain functional characteristics associated with agency such as goal-directed behavior, adaptability, and causal efficacy without possessing the intrinsic features required for genuine moral agency. These capacities are not self-originating but derivative, grounded in human design, training data, and institutional contexts. As John Searle argues, artifacts possess only derived intentionality, since their apparent purposiveness depends on human assignment and recognition (Searle, 1995, 44–47). AI extends this notion by operating in ways that simulate agency while remaining ontologically dependent on human systems of meaning.

This derived character becomes more evident when AI is situated within human–AI assemblages, where action emerges from the interaction of multiple elements rather than from a single autonomous source. In such assemblages, outcomes are co-produced by human decisions, algorithmic processes, and institutional frameworks. For example, an AI-assisted medical diagnosis depends on data curation, model design, clinical interpretation, and regulatory standards. Agency in this context is relational and distributed, not individual and self-contained. This aligns with contemporary approaches in AI ethics that emphasize distributed agency and responsibility across socio-technical systems (Floridi et al., 2018, 693).

The move toward a hybrid ontology also resonates with insights from Ludwig Wittgenstein and Robert Brandom. Wittgenstein's emphasis on forms of life suggests that meaning and action arise within shared practices rather than from isolated entities (Wittgenstein, 1953, §241). Similarly, Brandom's account of the space of reasons situates agency within discursive practices involving commitment and accountability (Brandom, 1994, 157–162). AI systems, while not full participants in these normative structures, become integral to the environments

in which such practices unfold. They shape the conditions of reasoning and decision-making without themselves entering the domain of normative agency.

This perspective enables a rethinking of agency beyond binary categories. Instead of asking whether AI is either an agent or an instrument, it is more accurate to distinguish between functional agency and normative agency. AI systems possess a form of functional agency, in that they can act within defined parameters, process information, and produce outcomes. However, they lack normative agency, which involves the capacity to justify actions, undertake commitments, and be held responsible within a moral framework (Brandom, 1994, 188–190). This distinction avoids both anthropomorphism and reductionism by recognizing the unique position of AI.

The ethical implications of this hybrid ontology are significant. In complex socio-technical systems, responsibility cannot be attributed solely to individual actors nor displaced onto machines. Instead, it must be understood as distributed across multiple levels, including designers, developers, organizations, and regulatory institutions. As Luciano Floridi argues, the ethical evaluation of AI systems requires attention to the entire network in which they operate, rather than focusing narrowly on the technology itself (Floridi et al., 2018, 690–692).

At the same time, recognizing distributed responsibility does not entail the erosion of accountability. On the contrary, it demands the development of robust governance frameworks that ensure transparency, traceability, and ethical oversight. Mechanisms such as algorithmic auditing, explainability, and institutional regulation become essential for maintaining accountability in AI-mediated environments. These measures help prevent accountability gaps, where responsibility becomes diffused due to the complexity of technological systems.

In conclusion, the hybrid ontology of AI offers a more adequate framework for understanding its ethical status. By conceptualizing AI as a derived and quasi-agentic entity embedded within relational networks, it becomes possible to move beyond simplistic dichotomies. This middle path preserves the centrality of human moral agency while acknowledging the transformative role of AI in reshaping the structure of action and responsibility in contemporary society (Searle, 1995, 94–96; Brandom, 1994, 157–162).

9. Ethical Implications and Contemporary Challenges

The rapid integration of Artificial Intelligence into social, economic, and political domains has generated a range of ethical challenges that extend beyond abstract philosophical debate. These challenges arise from the concrete ways in which AI systems shape human lives, often in opaque and unevenly distributed ways. Understanding the ethical implications of AI requires attention to issues of bias and fairness, accountability gaps, legal and social consequences, and the importance of incorporating global and non-Western perspectives.

One of the most widely discussed concerns is the problem of bias and algorithmic harm. AI systems are trained on large datasets that often reflect existing social inequalities and historical

patterns of discrimination. As a result, these systems can reproduce and even amplify biases related to race, gender, class, and other social categories. For example, facial recognition systems have been shown to perform less accurately on darker-skinned individuals, while algorithmic decision-making in hiring or lending may disadvantage certain groups. These outcomes are not merely technical errors but raise serious ethical questions about fairness, justice, and the distribution of harm. As Cathy O'Neil argues, such systems can function as “weapons of math destruction,” embedding inequality within seemingly objective processes (O'Neil, 2016, 3–8).

Closely related to bias is the challenge of ensuring fairness in AI systems. Fairness is not a single, universally agreed-upon concept but encompasses multiple, sometimes conflicting criteria, such as equality of outcomes, equality of opportunity, and procedural justice. Designing AI systems that satisfy these competing notions of fairness is a complex task that requires not only technical solutions but also ethical and political judgment. As Luciano Floridi and his collaborators emphasize, ethical AI must be grounded in principles such as transparency, accountability, and respect for human values (Floridi et al., 2018, 690–692).

Another major issue is the emergence of accountability gaps in AI systems. As AI becomes more autonomous and embedded in complex socio-technical networks, it becomes increasingly difficult to trace the origins of specific decisions or outcomes. When harm occurs, it may be unclear whether responsibility lies with developers, data providers, users, or institutions. This diffusion of responsibility creates what has been described as an “accountability gap,” where no single actor can be clearly held accountable. From a philosophical standpoint, this challenges traditional models of moral responsibility, which assume identifiable agents acting with intention and awareness. Addressing this issue requires rethinking responsibility as distributed across networks while ensuring that accountability is not lost in the process (Floridi et al., 2018, 691–693).

The legal and social consequences of AI further complicate these ethical concerns. Existing legal frameworks are often ill-equipped to deal with the unique features of AI, such as opacity, scalability, and autonomy. Questions arise regarding liability in cases involving autonomous vehicles, algorithmic decision-making in criminal justice, and the use of AI in surveillance. Should responsibility be assigned to the programmer, the company, the user, or some combination of these? Legal systems must adapt to these challenges by developing new categories and principles that reflect the realities of AI-mediated action.

Socially, the widespread adoption of AI has implications for employment, power distribution, and human autonomy. Automation may displace certain forms of labor while creating new inequalities between those who control AI technologies and those who are subject to them. Moreover, the increasing reliance on AI in decision-making processes may erode human judgment and reduce opportunities for meaningful participation. These developments

raise concerns about the future of human agency and the preservation of democratic values in technologically mediated societies.

Finally, it is essential to consider global and non-Western perspectives in AI ethics. Much of the current discourse is shaped by Western philosophical traditions and technological priorities, which may not adequately reflect the values and concerns of other cultural contexts. For instance, concepts of community, relationality, and harmony play a central role in many non-Western traditions, offering alternative ways of thinking about responsibility and ethical engagement. Incorporating these perspectives can enrich the ethical analysis of AI and promote more inclusive and context-sensitive approaches.

In particular, traditions such as Indian philosophy emphasize interconnectedness and duty (dharma), while African philosophies like Ubuntu highlight communal relationships and shared humanity. These frameworks challenge individualistic models of agency and responsibility, suggesting that ethical evaluation should focus on relationships and collective well-being rather than isolated actions. Engaging with such perspectives can help address the global impact of AI and ensure that ethical frameworks are responsive to diverse social realities.

In conclusion, the ethical implications of AI are multifaceted and deeply interconnected. Issues of bias, fairness, accountability, legal responsibility, and cultural diversity all point to the need for a comprehensive and inclusive approach to AI ethics. Rather than viewing these challenges in isolation, they must be understood as part of a broader transformation in the nature of agency and responsibility in the age of intelligent machines.

Conclusion

This study has examined the ethical ontology of Artificial Intelligence by critically analyzing whether AI should be understood as a moral agent or merely as an instrument. The discussion began by clarifying key conceptual distinctions, including the nature of agency, instrumentality, and moral responsibility. It then explored classical and contemporary ontological frameworks, showing how traditional views of humans as agents and machines as artifacts are increasingly challenged by the complexity of AI systems. The analysis of AI as instrumentality emphasized that AI functions as an extension of human intention, with responsibility ultimately grounded in human actors and institutions. In contrast, the examination of AI as agency demonstrated that current systems fail to meet the essential criteria for moral agency, such as intentionality, consciousness, and normativity.

Philosophical perspectives from John Searle, Ludwig Wittgenstein, and Robert Brandom further reinforced this conclusion by highlighting the importance of intrinsic intentionality, forms of life, and participation in the space of reasons conditions that AI does not satisfy. At the same time, the analysis recognized that AI cannot be reduced to a simple tool, as it plays an active role in shaping decisions and outcomes within socio-technical systems.

Accordingly, the paper has advanced a middle position: AI is best understood as occupying an intermediate space between agency and instrumentality. It functions as a derived or quasi-agent, embedded within networks of human and institutional practices, where agency and responsibility are distributed rather than isolated.

This conclusion has important implications for future research in philosophy and AI ethics. It calls for the development of more refined theories of distributed agency, as well as interdisciplinary approaches that integrate philosophy, technology, and social sciences. It also highlights the need for ethical frameworks that address accountability in complex systems while remaining sensitive to global and cultural diversity. Ultimately, understanding AI in this nuanced way can help guide responsible innovation and ensure that ethical responsibility remains firmly grounded in human practices.

References

- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.
- Chalmers, J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Coeckelbergh, M. (2020). *AI ethics*. MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Floridi, L. (2013). *The ethics of information*. Oxford University Press.
- Floridi, L., Cowls, J., Beltrametti, M., & et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gunkel, D. J. (2018). *Robot rights*. MIT Press.
- Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21. <https://doi.org/10.1109/MIS.2006.80>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. <https://doi.org/10.2307/2183914>
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Searle, J. R. (1995). *The construction of social reality*. Free Press.
- Searle, J. R. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.
- Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.