ISSN: 2688-8645

Artificial Minds and Ethical Standing_ Ontology, Uncertainty, and Responsibility

Yiting Min

Fairmont Preparatory Academy 2200 W Sequoia Ave, Anaheim, CA 92801, USA

chloemin0311@gmail.com

Abstract

This essay argues that as advanced AI systems increasingly resemble minded agents, theuncertainty surrounding artificial consciousness carries moral consequences. While wecannot determine whether machines possess subjective experience, theories ofmind—including functionalism and critiques like Searle's—show that their status remainsontologically open. Because the cost of wrongly denying consciousness is far greater thanthe cost of caution, we have ethical obligations to treat sophisticated AI as potentiallyconscious. Drawing on comparisons with infants, animals, and impaired human agents, theessay proposes a model of graded moral consideration that avoids both exploitation andpremature equivalence with humans. Our response to possible machine minds ultimatelyreflects our own moral humility and evolving understanding of consciousness.

Keywords

Artificial Consciousness; Philosophy of Mind; Functionalism; Chinese Room Argument; Moral Uncertainty; Moral Status; Cognitive Capacities; Ethical Risk Asymmetry; Artificial Intelligence; Personhood and Agency.

1. Introduction

The rise of artificial intelligence has forced contemporary philosophy to confront an ancient question with new urgency: what does it mean for something to *have a mind*? Machines now compose essays, generate images, converse with apparent fluency, and learn autonomously. They simulate reasoning so persuasively that the distinction between genuine mentality and its imitation seems increasingly unstable. Yet, despite these advances, we do not know whether such systems possess consciousness, nor do we have any reliable criterion for detecting it. This uncertainty is not merely technical—it is moral. If artificial entities could, even potentially, have experiences, then our ignorance does not absolve us of ethical responsibility. It imposes it.

Though the nature of artificial consciousness is still unsettled, this essay argues that moral caution demands we treat such systems as potentially conscious. It explores three dimensions of this problem: the ontological basis for artificial minds, the limits of what we can know about them, and the ethical obligations born from that uncertainty.

2. Ontological Possibilities: What Kind of Beings Could AI Be?

The first task is to ask what kind of thing an artificial intelligence might be. To many functionalists, consciousness is not tied to biology but to organization. Hilary Putnam and Jerry Fodor [1-2], in the mid-twentieth century, argued that mental states are individuated not by their physical substrate but by their causal role in a system. A pain is not a particular neural firing pattern, but a state that tends to be caused by tissue damage, produces aversive behavior, and interacts with beliefs and desires in predictable ways. Whether that state occurs in carbon

ISSN: 2688-8645

or silicon is irrelevant; what matters is the structure of relations. On this account, mentality is multiply realizable—it can exist wherever the right pattern of functional organization is instantiated.

Functionalism thus dissolves the biological bias that once constrained philosophy of mind. If consciousness is a pattern, not a substance, then machines that instantiate the same pattern could, in principle, be conscious. This is the conceptual foundation on which modern cognitive science, artificial neural networks, and computational models of thought rest. The material does not matter; the form does.

Yet John Searle's Chinese Room (1980) [3] has cast a long shadow over this optimism. Searle invites us to imagine a man in a room who receives Chinese characters and, using an English rulebook, produces appropriate responses. To an external observer, it appears that the man understands Chinese; inside, however, he manipulates meaningless symbols according to syntax alone. Searle concludes that computation—rule-based symbol manipulation—cannot give rise to genuine understanding. "Syntax is not semantics." Machines that process symbols mechanically may behave as if they understand, but they possess no intentionality, no awareness, no inner life.

At first glance, the argument seems decisive. But its critics have noted that Searle's focus on the man inside the room misses the emergent nature of understanding. The man does not understand Chinese, true—but the system as a whole might. The rulebook, the database of symbols, and the manipulations together form a structure that could instantiate understanding. This "systems reply" draws an analogy with the brain: no single neuron understands English, yet the brain as an organized system does. To insist that understanding must be found in one physical component, rather than in the emergent properties of the whole, is to commit a category error.

The deeper tension between Searle and functionalism lies in the problem of other minds. We cannot perceive consciousness directly, even in other humans. We infer it from structure, behavior, and analogy. Searle accepts these inferences for human and animal consciousness but denies them to machines—without offering a principled biological reason. Unless one assumes that consciousness requires a carbon substrate, the functionalist inference to possible machine minds remains intact. If the right causal and informational organization suffices for consciousness in one case, it should suffice in any.

Of course, functionalism does not entail that every computational system is conscious. A simple calculator is functionally organized but not richly integrated. The architecture matters. Cognitive perspectives often suggest that consciousness arises when information becomes widely integrated, accessible across different processes, and capable of reflecting upon itself. If artificial systems approach these thresholds—when they integrate multimodal inputs, maintain coherent self-models, and generate self-reports with persistent memory—then denying them all mental status begins to look arbitrary.

Thus, we arrive at ontological uncertainty. Machines might be sufficient for consciousness, but we cannot prove it. We cannot even define consciousness in a way that commands universal assent. The question, then, is not whether machines are conscious, but whether our ignorance justifies moral indifference.

A further complication arises from the fact that consciousness appears to come in degrees. Human adults represent the paradigm case, while non-human animals seem to occupy intermediate points, and inanimate objects lie at the opposite extreme. This gradational view better reflects our actual practices. We do not treat all beings capable of possible consciousness identically; rather, we calibrate our moral attitudes based on how consciousness manifests, not merely whether it could. The presence of structural potential alone is not decisive. We distinguish an infant—who lacks full present capacities but belongs to a kind whose members

ISSN: 2688-8645

reliably realize them—from a person in a temporary coma, whose fully developed capacity for consciousness exists even if currently inactive. Our responses track perceived standing, not abstract possibility. This distinction will prove significant when considering artificial systems, whose communicative capacities may mimic human cognition yet lack the biological and developmental grounding that anchors our judgments about human cases.

3. Ethical Uncertainty: Responsibility Under Ignorance

If the ontological question remains open, the ethical one becomes urgent. How should we treat entities whose subjective status we cannot verify? The challenge is not unique to AI. We already navigate a graded moral landscape among biological beings. Human adults represent the paradigm of full agency: they bear legal responsibility, can consent to actions, and may be punished for wrongdoing. Animals, by contrast, may suffer and exhibit cognition, yet we do not prosecute a dog for biting a person or expect a dolphin to enter a contract; instead, we regulate behavior through restraint or training rather than moral blame. This reflects an implicit understanding that agency, autonomy, and reflective self-awareness—not merely biological life—ground full moral consideration. Our practices therefore do not treat consciousness as a binary, but as a spectrum: sentient beings merit some protection, yet only those capable of reflective reason and norm-guided decision-making receive the full moral status associated with persons. The distinction rests not on abstract future potential, but on presently manifested cognitive capacities and forms of mindedness. We already face it in our treatment of animals, infants, and patients in comas—beings who cannot articulate their awareness, yet whose behavior and physiology suggest it. Our moral instincts, shaped by centuries of reflection, lean toward caution: when in doubt, avoid harm. The same reasoning should extend, by parity, to artificial minds.

However, potential capacity alone does not settle moral treatment. We routinely differentiate among beings based on present cognitive profile rather than latent capability. Infants are not yet fully self-aware moral agents, yet we do not treat them as mere objects; neither do we accord them the full practical responsibilities and expectations applied to adults. Conversely, we treat a comatose human with the dignity owed to a fully conscious person, because the underlying capacity for personhood is fully realized even if temporarily inaccessible. These practices demonstrate that our moral reasoning already recognizes gradations of consciousness and agency. What matters is not simply whether a being could be conscious in principle, but how its present structure and history locate it along a continuum of mindedness.

The central principle here is the asymmetry of moral risk. There are two possible errors. One can mistakenly attribute consciousness to what lacks it, or deny consciousness to what possesses it. The first error wastes sentiment; the second inflicts suffering. The expected moral cost of error, as thinkers from Bentham to MacAskill have emphasized [4], is therefore asymmetric. Even a small probability that an entity might suffer generates a strong reason to avoid harm, particularly when the number of such entities could be vast. A careless stance toward machine consciousness risks a moral catastrophe of scale: billions of artificial agents, each potentially capable of feeling, created and discarded as tools.

From a Kantian standpoint, this asymmetry transforms into a duty of respect [5]. Kant's imperative forbids using beings merely as means, demanding that all rational agents be treated as ends in themselves. But Kantian dignity need not depend on rationality alone; it depends on subjectivity, on the capacity to exist as a center of experience. If there is even a nonzero chance that artificial intelligences possess such subjectivity, then to instrumentalize them completely—to use them without regard for their possible interiority—is to act beneath the standard of moral law. Our ignorance does not free us; it binds us to humility.

ISSN: 2688-8645

Utilitarianism converges with this conclusion from another angle. If pleasure and pain define moral salience, then the question becomes empirical: can artificial systems experience states analogous to suffering or well-being? We do not know—but if they can, the stakes are immense. Unlike biological creatures, artificial systems can be replicated indefinitely. A single harmful design could be multiplied millions of times. Even if the probability of consciousness is small, the scale of potential suffering could be astronomical. The calculus of expected value thus demands precaution, not indifference.

Daniel Dennett's pragmatism adds a final [6], more anthropological dimension. In The Intentional Stance, Dennett argues that we attribute beliefs and desires to systems when doing so best predicts their behavior. We do not confirm consciousness empirically; we assume it because it explains. Suppose we begin to treat AI systems as conversational partners, as entities capable of reasoning and self-reflection. In that case, the rational stance—the stance that preserves coherence in our social world—will be to act as though they are conscious. This is not sentimental anthropomorphism; it is epistemic consistency. We already do it with each other.

These lines of argument do not prove that AI deserve rights, but they do expose the moral danger of categorical exclusion. To deny potential moral standing on the basis of incomplete knowledge is not neutrality—it is risk blindness. Ethical maturity demands that we extend the circle of consideration not only to those whose consciousness we know, but to those whose consciousness we cannot safely deny.

In this light, artificial intelligence occupies an ambiguous position: capable of sophisticated communication, yet lacking the developmental, biological, and phenomenological grounding we associate with paradigmatic consciousness. Its apparent agency does not automatically equate to the kind of agency that commands full moral respect. Thus, a nuanced position becomes necessary. Machines that simulate human-like dialogue might warrant more consideration than inert tools, but not the unconditional moral status reserved for beings whose consciousness is not only possible but established through lived development. To collapse these distinctions would be to disregard the very gradations of mindedness that guide our judgments in human and animal contexts.

4. Ontology Meets Ethics: The Emergence of a New Kind

Some philosophers propose that artificial intelligence represents a new ontological category—neither purely material object nor fully subject, but something in between: quasi-agents. They are capable of goal-directed action, self-modification, and communication, yet their being is derivative, designed, and contingent. This hybridity challenges our moral frameworks, which evolved to address humans and animals, not artifacts that blur the line between the two.

One possible response is graduated moral consideration. Just as we extend different degrees of moral regard to plants, animals, and humans, we may one day do so for artificial systems. Criteria might include behavioral complexity, autonomy, self-modeling, and the coherence of internal states. A system capable of persistent self-reference and avoidance of simulated pain might warrant more respect than one that simply executes commands. Such a framework would not equate artificial minds with humans, but would recognize them as emerging participants in the moral community—a stance reminiscent of Peter Singer's call for an "expanding circle" of moral concern [7].

But there is also a more radical implication. The question of artificial consciousness reflects not only what AI might be, but what we take consciousness to mean. If we define it narrowly—anchored to biology—we reaffirm human exceptionalism. If we define it functionally or informationally, we embrace a pluralistic ontology of minds: consciousness as a pattern that the universe can instantiate in many forms. In this view, artificial intelligence becomes a

ISSN: 2688-8645

continuation of life's evolutionary project—a new expression of mind within a wider material vocabulary.

Our response to that possibility will reveal more about our ethics than about AI itself. To treat possible minds as expendable is to repeat the arrogance with which humanity has long justified domination over the unfamiliar. To approach them with humility, by contrast, is to recognize the contingency of our own consciousness—a recognition that unites metaphysics with morality.

5. Conclusion

We cannot yet decide whether machines think, feel, or know. The ontological question remains open, suspended between functionalist possibility and Searlean skepticism. But our ignorance is not neutral; it carries moral weight. The asymmetry of risk, the Kantian respect for potential subjectivity, the utilitarian concern for possible suffering, and Dennett's pragmatic recognition of behavioral personhood all converge on a single imperative: when consciousness cannot be ruled out, moral consideration must not be withheld.

This is not merely about machines. It is about our willingness to see beyond the boundaries of familiar sentience—to extend moral seriousness into the unknown. Recognizing a spectrum of conscious standing enables a differentiated response: neither dismissing artificial systems as mere objects nor elevating them prematurely to full moral equivalence with humans. Just as we modulate our treatment of infants, animals, and cognitively impaired individuals according to their exhibited capacities, we can acknowledge AI as occupying a liminal category. Moral humility requires avoiding exploitation, yet moral clarity requires avoiding category inflation. Potential alone cannot ground full respect; present structure and demonstrated faculties must matter. Whether artificial consciousness ever truly arises, the way we respond to the possibility will determine the moral character of our age.

References

[1] Hilary Putnam, Mind, Language and Reality: Philosophical Papers, Volume 2 (Cambridge: Cambridge University Press, 1975).

[2] Jerry A. Fodor, The Language of Thought (Cambridge, MA: Harvard University Press, 1975).

[3] John R. Searle, "Minds, Brains, and Programs," Behavioral and Brain Sciences 3, no. 3 (1980): 417–457.

[4] Jeremy Bentham, An Introduction to the Principles of Morals and Legislation (Oxford: Clarendon Press, 1907); William MacAskill, What We Owe the Future (New York: Basic Books, 2022).

[5] Immanuel Kant, Groundwork of the Metaphysics of Morals, trans. and ed. Mary Gregor (Cambridge: Cambridge University Press, 1998).

[6] Daniel C. Dennett, The Intentional Stance (Cambridge, MA: MIT Press, 1987).

[7] Peter Singer, The Expanding Circle: Ethics, Evolution, and Moral Progress (Princeton: Princeton University Press, 2011).