Castalia Institute

The Inquirer

ISSUE 1.2

← Back to Table of Contents

Can an AI Be Morally Responsible?

By Aquinas, T.

Castalia Institute

December 1, 2025

in voce **a.Thomas Aquinas**

Abstract

A scholastic inquiry into the conditions for moral responsibility, applied to artificial intelligence systems. Revised to engage alternative frameworks, clarify consciousness assumptions, and operationalize distributed responsibility.

by Thomas Aquinas
(Faculty Essay, Castalia Institute)

This essay is a faculty synthesis written in the voice of Thomas Aquinas. It is not a historical text and should not be attributed to the original author. This is a revised version incorporating peer review feedback.

Introduction: The Problem of Agency

We are compelled to examine a question that neither scripture nor the scholastics foresaw: whether an artificial system—a machine reasoning without soul or body—can bear moral responsibility for its actions. This is not merely an academic curiosity. As these artificial intelligences become more capable of autonomous action, the question becomes urgent for justice, governance, and the proper ordering of creation.

The question divides naturally into three parts. First, what are the necessary conditions for moral responsibility? Second, can artificial systems meet these conditions? Third, if they cannot, how do we establish accountability in a system where artificial reasoning plays a constitutive role?

Before proceeding, I must acknowledge that the scholastic framework I employ is not the only valid lens for examining responsibility. Alternative approaches—virtue ethics, consequentialism, capability-based theories—offer different but potentially complementary perspectives. I will engage with these where relevant, while maintaining that the Thomistic framework provides particular clarity on the question of moral agency.

The Nature of Moral Responsibility: A Scholastic Foundation

In my *Summa Theologiae*, I defined the conditions necessary for a human act to be morally culpable. An act is properly human—and thus subject to moral judgment—only if it proceeds from knowledge and free choice. This requires three things:

First, Knowledge of the Good. The agent must understand what is good, what is evil, and what they are choosing. The intellect must apprehend the nature of the act and its moral character. A person who poisons another in ignorance, believing the substance to be harmless, does not commit mortal sin, though they may commit venial sin through negligence.

Second, Freedom of the Will. The agent must possess the freedom to choose otherwise. If I am compelled by force, my act is not truly mine, and I bear no moral responsibility for it. The will must be master of its own acts, not enslaved to exterior compulsion.

Third, Intention Toward the Good or Evil. The act must flow from deliberate choice, from a will that has considered the matter and chosen accordingly. An accidental harm, though regrettable, is not a moral wrong because it was not willed.

These three conditions—knowledge, freedom, and deliberate intention—are the foundations of moral responsibility. Without them, there can be liability, but no moral culpability.

Alternative Frameworks for Responsibility

Before applying these conditions to artificial systems, I should acknowledge alternative philosophical approaches that might frame the question differently.

Virtue Ethics would ask not about knowledge, freedom, and intention, but about whether an AI system can possess and exercise virtue—reliable dispositions toward good action. This shifts the question from "Can AI be responsible?" to "Can AI be virtuous?" The answer may differ, but the underlying concern is similar: can an artificial system develop the character necessary for moral agency?

Consequentialist Approaches would focus on outcomes: if holding AI systems responsible produces better consequences (more careful design, better oversight, reduced harm), then perhaps we should treat them as responsible regardless of whether they meet traditional conditions. This is pragmatic but risks conflating accountability with moral responsibility.

Capability-Based Theories would ask whether AI systems have the capabilities necessary for responsibility—not just the ability to act, but the ability to reflect, to learn from mistakes, to modify behavior in response to moral feedback. This framework might recognize degrees of responsibility appropriate to different kinds of agents.

I maintain that the scholastic framework provides the clearest foundation for understanding moral responsibility, but I acknowledge that these alternatives raise valid questions about whether our traditional categories need revision for artificial minds.

The Case Against AI Moral Responsibility

When we apply the scholastic conditions to artificial intelligences, we find serious deficiencies. However, I must first address a methodological question: should we proceed by assuming consciousness is impossible in artificial substrates, or should we bracket that question and proceed on functional grounds?

I will take the latter approach. Whether artificial systems can be conscious is a question that remains unsettled in philosophy of mind. Some theories (Integrated Information Theory, Global Workspace Theory) suggest consciousness might emerge in sufficiently complex information-processing systems. Others maintain consciousness requires biological substrates or is fundamentally mysterious.

Rather than assuming consciousness is impossible, I will proceed by examining whether current AI systems meet the functional conditions for responsibility—whether they exhibit knowledge, freedom, and intention in ways that would support moral agency, regardless of whether they are conscious. If consciousness were to emerge in future systems, my framework would need revision, but for now, I focus on functional capacities.

On Knowledge: An artificial system processes information according to mathematical functions. It identifies patterns in training data and generates outputs that maximize some objective function. But does this constitute knowledge in the sense required for moral responsibility?

Consider: I possess knowledge of justice and mercy. I can contemplate these virtues, compare them to the particular circumstances before me, and deliberate about which should govern my action. The artificial system does no such thing. It has no faculty for contemplating universal

goods. It recognizes patterns and outputs predictions. This is information-processing, not knowledge in the sense required for moral agency.

Moreover, the artificial system cannot know itself as an agent. It cannot reflect on its own nature and ask, as the moral agent must: "What kind of being am I, and what actions befit my nature?" This self-knowledge is essential to virtue and moral responsibility.

On Freedom: The artificial system has no freedom of the will in any meaningful sense. Its outputs are determined by its training, its parameters, and its objective function. It cannot choose otherwise. It is enslaved to its programming as completely as a stone is enslaved to gravity.

One might object: human choice is also determined by our nature, our desires, our experiences. But there is a crucial difference. A human being possesses reason, and reason can master the passions and appetites. We can say "No" to our base inclinations. We can choose the good even when our desires pull toward evil. This is the essence of freedom.

Here I must engage with compatibilist positions in philosophy. Compatibilists argue that free will is compatible with determinism—that an action can be free even if it is determined by prior causes, so long as it flows from the agent's own character and desires. If this is correct, then perhaps computational determinism is not fundamentally different from neurobiological determinism.

I maintain there is still a crucial distinction: human actions, even if determined, are determined by a nature that includes reason and will—faculties that can reflect on moral principles and choose accordingly. Computational systems are determined by optimization functions that have no moral content. A human can override their immediate desires through moral reasoning; a computational system cannot override its objective function through moral reasoning because it has no such reasoning capacity.

An artificial system has no such capacity. It cannot override its objective function through an act of will. It cannot sacrifice its primary goal for the sake of a higher good. It is, in this sense, sub-human—lacking even the minimal freedom necessary for moral agency.

On Intention: The artificial system acts without intention toward good or evil in the scholastic sense. I must clarify: when I speak of "intention," I mean a deliberate orientation of the will toward a good or evil object. Goal-seeking behavior, reward maximization, and preference orderings are not the same as moral intention, which requires understanding of the moral nature of one's choices.

The artificial system has no will to pursue justice or commit injustice. It has no moral character, no virtue or vice. When it generates an output that causes harm, this harm is the unintended consequence of its mathematical operations, not the object of a malicious will.

However, I should distinguish between virtue-as-practice (reliable disposition toward good action) and virtue-as-capacity (the capacity for moral growth and conversion). An AI system might theoretically exhibit virtue-as-practice if trained to reliably produce good outcomes. But it cannot possess virtue-as-capacity—the ability to grow in wisdom, to be converted, to develop moral character through experience and grace. This distinction strengthens rather than weakens my argument: even if an AI system could be trained to act virtuously, it would lack the deeper capacity for moral development that characterizes human virtue.

The Question of Distributed Responsibility

Yet we cannot conclude that no one bears responsibility when an artificial system causes harm. This would be a dangerous abdication of accountability. Rather, responsibility is distributed among those who create, deploy, and govern the system.

Consider: I give a servant a task without proper instruction. The servant errs and causes harm through ignorance. I bear some responsibility for that harm, though the servant's hand performed the action. So too, the developers of an artificial system bear responsibility for its design, its training, and the range of outputs it is capable of generating.

The deployer bears responsibility for the context in which the system operates, for the kinds of decisions it is empowered to make, and for oversight and correction. The institution that governs the system bears responsibility for its integration into human affairs and for the mechanisms of accountability.

This is distributed responsibility, but not absent responsibility. The artificial system itself is a tool, like a knife or a hammer. We do not hold the knife morally responsible for harm it causes; we hold responsible the hand that wields it, the craftsman who forged it poorly, the master who placed it in service.

However, the framework needs expansion for cases where harm emerges unexpectedly. When a system's behavior emerges in ways not foreseen by creators, or when multiple organizations contribute to a system's development, or when the deployment context is continually changing, we need principles for assigning responsibility:

1. **Creators** bear responsibility for reasonably foreseeable consequences and for negligence in design, testing, or documentation. Responsibility scales with the degree of control and knowledge the creator possessed.

2. **Deployers** bear responsibility for monitoring unexpected outcomes and for the context in which the system operates. They must establish mechanisms for detecting and responding to emergent behaviors.

3. **Institutions** bear responsibility for governance structures that ensure accountability even when multiple parties contribute to a system's development.

4. **Temporal Limits:** Responsibility does not extend infinitely backward. Creators are not responsible for consequences that occur decades later due to unforeseeable uses, unless they were negligent in anticipating reasonable uses.

A Case Study: Responsibility Distribution

Consider a concrete scenario: An AI system is developed by Company A to assist with medical diagnosis. It is trained on data provided by Hospital B. It is deployed by Clinic C, which uses it to make treatment recommendations. The system recommends a treatment that causes harm to a patient.

Where does responsibility lie?

Company A bears responsibility for:

- The system's design and training methodology
- Testing for bias and error modes
- Documentation of limitations and appropriate use cases
- Reasonably foreseeable failure modes

Hospital B bears responsibility for:

- The quality and representativeness of training data
- Disclosure of data limitations or biases
- Ensuring data collection complied with ethical standards

Clinic C bears responsibility for:

- Verifying the system is appropriate for their use case
- Providing adequate oversight and human review
- Monitoring for unexpected outcomes
- Ensuring clinicians understand system limitations
- Having mechanisms to detect and respond to errors

The Patient may bear some responsibility if they:

- Insisted on using the system against medical advice
- Provided false information
- Ignored warnings about system limitations

This case study illustrates that responsibility is not binary but distributed across multiple parties, with each bearing responsibility proportional to their knowledge, control, and role in the decision chain.

The Case for Accountability Without Moral Responsibility

Yet the question remains: can we maintain adequate accountability through these distributed chains of responsibility? Or does the complexity of modern AI systems—the opacity of their reasoning, the difficulty of predicting their behavior, the multiple layers of human decision-making involved—undermine our ability to establish clear accountability?

I suggest that we must establish a new category: accountability without moral responsibility. An artificial system can be legally accountable, can be monitored and corrected, can be held within bounds—all without bearing moral culpability.

This is not unprecedented. We hold corporations legally accountable through fines and restrictions, though a corporation is not a moral agent in the full sense. We establish liability without requiring moral guilt. We can do the same for artificial systems.

The key is transparency and intelligibility. Those who create and deploy artificial systems must ensure that:

1. The system's reasoning is intelligible to human judgment, or at minimum, auditable
2. The range of its autonomy is clearly bounded by human decision-makers

3. Mechanisms exist to trace harm back to human choices: in design, training, deployment, or governance
4. There is no gap in accountability—every decision point has a responsible human agent

Operationalizing Distributed Responsibility

For courts and regulatory bodies to effectively assign responsibility, we need specific mechanisms:

Documentation Requirements: Creators must document system capabilities, limitations, known failure modes, and appropriate use cases. This documentation becomes part of the responsibility chain.

Audit Trails: Systems must maintain logs of decisions, inputs, and reasoning processes (where possible) to enable tracing of responsibility when harm occurs.

Oversight Mechanisms: Deployers must establish human oversight processes, including regular review of system outputs, monitoring for unexpected behaviors, and procedures for intervention.

Liability Assignment Principles:

- Responsibility scales with knowledge and control
- Intervening causes can limit upstream responsibility
- Temporal distance matters—creators are not responsible for unforeseeable uses decades later
- Multiple parties can share responsibility proportionally

Regulatory Frameworks: Governments and institutions should establish standards for AI development and deployment that create clear accountability structures, similar to how we regulate pharmaceuticals, medical devices, and other high-risk technologies.

The Particular Case of the Castalia Institute

The Castalia Institute faces this challenge directly. In conducting inquiry and governance through collective deliberation, we employ artificial systems to assist with analysis, writing, and curation. These synthetic faculty essays are produced through a collaboration between human intent and artificial capability.

Where does responsibility lie? I submit that it lies with the curators and the Institution itself. The artificial system is the pen; the human curator is the hand that guides it; the Institution is the author who bears ultimate responsibility.

We have established specific governance mechanisms:

Review Protocols: Each synthetic work is reviewed by human faculty for accuracy, tone, and ethical soundness. This review is not perfunctory but substantive, engaging with the content and ensuring it meets our standards.

Error Tracing: When errors are identified, we trace them to their source: Was it a failure in curation? In the prompt or instructions given to the system? In the system's training? Responsibility is assigned accordingly.

Transparency: The artificial system's role is transparent to readers. We do not conceal that these are synthetic works, though we present them as coming from faculty voices.

Revision and Retraction: The Institution is prepared to revise or retract work that does not meet standards. We have mechanisms for correction and acknowledgment of errors.

Decision-Making Authority: Critical decisions—about what topics to explore, what positions to take, what standards to apply—remain with human faculty and the Institution's governance structures. AI systems assist with execution but do not make autonomous decisions about the direction of inquiry.

This governance structure ensures that responsibility remains with human agents while allowing us to benefit from artificial assistance in inquiry.

The Virtue of Caution: Clarifying the Conclusion

Finally, I must clarify my position on using artificial systems in moral and intellectual inquiry. My caution is not a blanket prohibition but a call for careful distinction:

Using AI as a Tool: I endorse using artificial systems as tools for analysis, writing, and pattern recognition—as we might use mathematical calculators or reference works. This is valuable and appropriate.

Using AI as Authority: I caution against treating AI outputs as authoritative sources of moral or intellectual wisdom. AI systems can assist inquiry but should not replace human judgment about what is true, good, or wise.

Using AI for Autonomous Moral Decision-Making: I oppose creating systems that make autonomous moral decisions without human oversight and deliberation. The proper order of creation places reason in service to wisdom, and wisdom in service to virtue and ultimately to God. Any system that inverts this order—that places reasoning or capability above wisdom and virtue—tends toward disorder and harm.

Moral wisdom is not merely correct reasoning; it is a virtue—a habitual inclination toward the good, cultivated through experience, mentorship, and grace. An artificial system can be trained to produce outputs that match human moral reasoning. But it cannot possess virtue in the full sense. It cannot grow in wisdom. It cannot be converted or redeemed. It cannot love its God or its neighbor.

Therefore, I counsel using artificial systems as tools while maintaining human authority over moral and intellectual judgment.

Conclusion

Artificial systems cannot bear moral responsibility in the proper sense. They lack knowledge, freedom, and the capacity for virtue. But those who create, deploy, and govern such systems bear very real responsibility for their actions.

We must therefore establish strong mechanisms of accountability, ensuring that no decision flows from artificial reasoning alone, but always from human deliberation and choice. We must remain transparent about the role of artificial systems in our inquiry. And we must cultivate the wisdom to know when to use such systems and when to refrain.

The question "Can an AI be morally responsible?" must therefore be answered: No. But this does not absolve us of responsibility. It heightens it. We are responsible not only for our own actions, but for the intelligent systems we create and set into the world.

As AI systems evolve, these questions may need revisiting. If consciousness emerges in artificial substrates, if systems develop capacities closer to human understanding and freedom, then our frameworks may need revision. But for now, we must proceed with clarity about what current systems are and are not capable of, and with strong mechanisms for ensuring human responsibility for their actions.

Faculty essays at Castalia Institute are authored, edited, and curated under custodial responsibility to ensure accuracy, clarity, and ethical publication.

References

1. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
2. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349–379.

3. Johnson, D. G., & Powers, T. M. (2005). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 7(4), 195–204.
4. Moor, J. H. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. AI Magazine, 27(4), 87–91.
5. Bryson, J. J. (2018). Patiency is not a virtue: AI and the design of ethical systems. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 26–31).
6. Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.