Journal of Philosophical Theological Research
Print: 1735-9791
Online: 2538-2500

# Agency Attributions and AI Agency

**Hajj Muhammad Legenhausen**

Professor of Philosophy, The Imam Khomeini Education and Research Institute, Qom, Iran.

legenhausen@gmail.com

## Abstract

The contemporary debate surrounding AI agency is frequently stalled by an essentialist preoccupation: attempting to discern whether an artificial system possesses genuine autonomy or traditional belief-desire agency. This paper argues that such an approach is a category error that obscures the practical realities of AI deployment. By shifting the discourse from an ontological investigation of machine agency to an "attributionist" framework, we can better understand how agency is constructed through socio-technical relationships rather than residing solely in algorithmic code. Drawing upon the theological concept of *fā'il bi-l-tajallī* – agency expressed through mediated manifestation – this paper posits that AI systems act as conduits for human intent and institutional structure. This perspective provides a necessary alternative to current technical debates. Furthermore, by engaging with recent scholarship on responsibility gaps – specifically by synthesizing Sven Nyholm's seminal identification of the moral vacuum in autonomous systems with Herman Veluwenkamp's granular analytic distinctions – we move toward a more precise taxonomy of accountability. Veluwenkamp shows that the so-called "responsibility gap" must be disaggregated into Epistemic Responsibility Gaps (ERG), relating to system complexity and explainability, and Control Misalignments (CM), which stem from institutional delegation failures. Ultimately, this paper reframes the "Alignment Problem." It is not a purely technical challenge of matching an objective function to a static goal, but a socio-technical endeavor of normative integration. An aligned AI is not one that achieves perfect independence, but one whose operations remain within the legible, revisable, and accountable bounds of human institutional intent. Through this attributionist lens, we demonstrate that resolving alignment concerns requires less focus on the internal "will" of the machine and greater focus on the distributed responsibilities that arise with the instantiation of these systems.

Original Research

## Keywords:

agency, AI agent, reflex, action, autonomy, responsibility gap, agent by self-disclosure

**Received:** 2026/02/16; **Received in revised form:** 2026/04/12; **Accepted:** 2026/05/20; **Published online:** 2026/05/25

Legenhausen, H. M. (2026). Agency Attributions and AI Agency. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 77-90. https://doi.org/10.22091/jptr.2026.17198.3657

Copyright © The Author

## Introduction: What Is Agency?

Agency is the ability to perform actions. Being an agent is having this ability. Agency is essential to personhood. Persons do things. The converse is not the case. There is agency in some animals we do not count as persons. This approach pushes the question of agency to inquiries into what counts as an action. In contemporary action theory, it is usually assumed that actions must be intentional. [1] Intentionality distinguishes actions from involuntary movements. In Islamic philosophy, however, intentional agency is only one kind of agency. God is considered to be an agent by self-disclosure, *fāʿil bi-l-tajallī*, and not an intentional agent. [2] Here, I will not assume that actions have to be intentional, even if it is granted that human action normally involves intentionality. Clearly, actions need not be intentional for at least some non-human agents.

It is widely assumed that human actions are bodily movements that are caused by the right combination of beliefs and desires. This approach to action and agency has a number of drawbacks (aside from making divine agency incomprehensible, given a classical view of theism according to which God has no desires). It also makes AI agency impossible, given that AI agents do not have bodily movements. As an alternative, I will focus on the issue of agency attributions. An action must be performed by an agent in such a manner that the action can be attributed to the agent as something the agent did. By considering the conditions under which the performance of an action is attributed to an agent, we may outline several kinds of agency attribution, types of agency that are attributed, and consider the variety of things to which agency is attributed, including God, angels, humans, some animals, and some AI systems. There is also social agency that can be applied to groups of animals who form swarms, herds, and packs to protect their members and hunt. Human corporate agents include tribes, states, commercial conglomerates, orchestras, political movements, and governmental agencies. With respect to responsibility, AI agents may be more similar to corporate than individual agents, since responsibility for what AI agents do devolves to corporations, designers, developers, and users, as well as regulatory agencies.

There are a number of philosophical issues that I will not take up here. There is a debate about whether corporate actions are reducible to individual actions. I will not assume they are, but I will not try to argue that they are not. Likewise, I will be non-committal about whether agency reduces to justified agency attributions or to socially accepted agency attributions, and I will take

1. For an influential critical discussion of intention, see Anscombe (1959).
2. I discuss this in Legenhausen (2021).

no position on the relationship between agency and freedom of the will. [1]

By utilizing agency attributions instead of relying directly on a theory of the nature of agency, we will rely on scalar attributions, since those making the attributions may be more or less inclined to ascribe a given kind of agency to a system. Instead of imagining agency as a property that a system either has or lacks, consider a weighted set of capacities and relations due to which various kinds of agency attributions will be made. Different "agency markers" (goal-directedness, adaptivity, initiative, normative sensitivity, etc.) can vary independently, yielding different levels or types of agency (Zhu et al., 2026).

The question of whether some AI systems have agency turns out to be misplaced. It would be more appropriate to ask what kinds of agency are attributed to an AI system, and by whom. All the parties that use AI are called stakeholders. If our task is to understand the senses in which various kinds of agency are attributed to AI systems by various groups of stakeholders, we should avoid a philosophical characterization of agency that makes it true by definition that AI is not capable of any sort of agency. At the same time, we should not claim that AI has agency just because of the widespread usage of the term "AI agent." There are several factors that can serve as reasons for agency attributions, and these may be weighted and applied in different ways by different groups of stakeholders.

Instead of beginning with a theory about the nature of agency, we might reject the question "What is agency?" and replace it with several others:

- What are the types of agency that can be attributed to AI systems?
- What kinds of evidence justify attributions of the various types of agency?
- What is the range of AI capabilities within which agency attributions are made?
- How do the attributions of agency to AI systems compare with attributions of agency to other entities?

## Actions and Reflexes

If agency is the ability to perform actions, as opposed to the ability to move, the difference between an action and a motion will aid our understanding of agency. In animals, the difference appears in how the animal is related to the movements of its body. According to Michael Tomasello's recent study of the

1. For a review of these topics see part 6 of Ferrero (2022).

emergence of agency in the course of biological evolution (Tomasello, 2022), the movement of an animal is an action when it is under the animal's control. Tomasello considers the question of how we can distinguish when a movement occurs as the result of some mechanism and when it is a controlled action. His analysis draws on cognitive psychology, which is not an especially promising route if we want to understand AI agency and AI agents do not possess psychological states. Nevertheless, we can transpose some of his claims from psychology to the behavioral observations that can support agency attribution.

Tomasello describes agency as a certain kind of capacity to behave in order to achieve goals. Not any sort of goal-oriented activity will be due to agency. A standard question in action theory is how to distinguish a self-regulating mechanism, like a thermostat, from an agent. In both cases, we need to be able to identify goal-directed activity. A goal attribution will be justified to the extent that it helps explain a pattern of behavior across circumstances – including adjustments, failures, and changes in the environment. It should also support some counterfactual expectations: if relevant circumstances had been different, we would expect the system to have adjusted its behavior in identifiable ways. This still is not sufficient to distinguish an agent from an HVAC (heating, ventilating, and air-conditioning) system controlled by a thermostat. We do not consider the HVAC system an agent because we can explain how the system works without any teleology. When the temperature gets too low or too high, the heat or the air conditioning turns on. The result is keeping the temperature of a room within a certain range. That is the goal of those who use the system to stay comfortable. It is not the goal of the thermostat (Tomasello, 2022, pp. 14–16). We shift from mechanical explanations to teleological ones when the mechanical explanations become so complex that they prevent human understanding of how the system works. Then we explain the behavior of a system as agential when it is flexibly directed toward goals through the utilization of feedback that enables the system to take a course that avoids obstacles and increases the likelihood of success in achieving a goal. We attribute agency when we observe behavior that requires this sort of explanation. The explanation through goals and feedback does not mean that there is no causal sequence of events that brought about the observed behavior. However, the causal sequence at the micro-level fails to enable us to understand what the animal is doing when agency is involved. If an animal twitches as a reflex to some stimulation, we can explain the twitch as a nervous reaction. When the animal uses skills to catch prey, nervous reactions are also in play; but the reactions cannot explain what is being done by the animal when goals, skills, threats, and opportunities are left out of the explanations. It is the kind of explanation needed to achieve understanding of behavior that justifies agency attributions, independent of whatever the phenomenology of agency

might involve in the agents. We do not have to know what it is like to be a bat in order to know what sorts of things bats do, and how these things are different from mere bat reflexes.

One might attempt to distinguish actions from reflexes by their relation to the intellect. Actions result from practical reasoning. Reflexes do not. While Aristotle held that animals do not possess intellect, he did admit that practical wisdom is attributed to them:

> … even some of the wild beasts are said to be practically-wise – those that appear to have a capacity for forethought about their life. (Aristotle, trans. 2014, 1141a).

Aristotle is not attributing agency and practical wisdom to the animals, because he denies that animals have intellect. However, he admits that the forethought exhibited by animals has led some people to attribute practical wisdom to them.

The distinction between voluntary and involuntary movements was recognized by Aristotle and in other ancient sources. Many centuries later, Ibn Sina, in the second book of his *Canon of Medicine*, describes "animal" functions of the body that facilitate immediate responses to external stimuli, and distinguishes between movements directed by the conscious intellect and those triggered automatically by sensory-motor pathways (Ibn Sina, 1999). Six centuries later, Descartes, in his treatise *L'Homme*, described ocular fixation and other bodily motions as the result of "animal spirits" flowing from the pineal gland through the nerves to muscles, with sensory traction on nerve filaments opening pores in the ventricles and thereby releasing more spirits – an automatic chain that can reinforce itself. It would take a couple more centuries for the physiological description of the "reflex arc" to be developed, although Descartes' German and English translators have interpreted his observations as indicating something like the modern concepts of reflex and feedback (Descartes, 1972, p. 98, n. 146).

Marshall Hall introduced the term "reflex arc" in 1833 for the path from stimulus through the nerves to movements in glands or muscles to produce a response; and he demonstrated that certain responses in decapitated animals could occur without any involvement of the brain. In the 1860s, Ivan Sechenov published studies of the nervous system in which self-regulation and the inhibition of reflexes were described.

In the 20th century, Norbert Wiener made the mathematical analysis of feedback loops a centerpiece of his cybernetics, the study of communication and control in living organisms or machines; and he illustrated this in an

explanation of ocular reflexes (Wiener, 2019, pp. 184ff.). Self-regulation through feedback loops would later become a key factor in the development of machine learning and artificial intelligence.

In physiology, the study of reflexes has continued to progress by differentiating the types of reflexes and the number of synapses involved, as well as the role of feedback loops and positive and negative reinforcement. Some reflexes, like breathing, can be brought under control. Even the pulse may be lowered through meditation. Agency is not attributed to reflexes when relatively limited components of the nervous system are involved.

For AI systems, we can find outputs that are generated by the system in an automatic way, like reflexes. AI reflexes can be contrasted with outputs that result from more complex processing that is designed to maximize expected success in a project set for the system by its users.

## AI Systems and Agency Attributions

There is a formidable number of technical terms used to describe AI systems that draw on the language of agency: there are AI agents, agentic systems, and various levels of AI autonomy. These terms are defined in the literature of the science of AI. This complicates our method of trying to understand AI agency through agency attributions. When experts and users use the language of agency to describe a system, they need not be attributing the same thing to it by calling it an agent. This poses a serious challenge to the approach advocated here of trying to understand AI agency by examining relevant agency attributions, because the mere use of words like "agency," "agent," and "agentic" is not sufficient to indicate that it is agency that is being attributed. There is no philosophically or biologically useful sense of agency that could be applied to a clock just because we say that it does something, namely, that it tells us the time. The difficulty can be overcome by recognizing that different groups of stakeholders may use the terminology of agency in different ways. We should not assume a core meaning of agency that can be discovered merely by examining attribution patterns.

There are a number of different groups that may attribute agency to some AI systems: academics in the field of AI, developers, users, courts, oversight agencies, and corporations that buy or sell AI services. If we are concerned with the systems themselves rather than how we interact with them as users, algorithmic complexity might provide a more useful account of grades of agency than the reactive attitudes we direct toward persons as responsible agents and withhold from very young children and those who suffer from

debilitating mental illnesses or who cannot be considered to have the competence needed for responsibility because of trauma or other reasons (Strawson, 2008; Bagnoli, 2021). The kinds of social legibility that prompt reactive attitudes, such as blame, resentment, praise, and gratitude, will be more pertinent to user attributions of agency to AI systems than to system developers. AI assistants, companions, and other kinds of systems can give users the impression of agency, as if the systems had a will of their own, because the responses they provide are unpredictable while remaining relevant. Marketing also encourages users to interact with the systems as if they were persons. The result is not only that users attribute human-like agency to the systems; the users' own agency is extended through devices to the systems that enable them to do things that they otherwise could not. The downside of this is that over-reliance on AI systems may have deleterious effects on the ability to make informed choices and may debilitate the user's own agency.

For developers and designers, the term "AI agent" is reserved for systems that have abilities that go beyond those of a simple AI assistant. While an assistant needs continual interaction with a user, an AI agent can complete tasks on behalf of users or other systems without further human inputs by planning its own workflow and using available tools. Systems are said to be autonomous to the extent that they can perform tasks without human intervention. Agency is ascribed to AI systems by considering the limitations of the information and tools that can be accessed in order to produce responses to inputs from the environment or from users, to perform specific tasks, or to carry out larger projects. In AI theory, systems are described that represent inputs and compress information. Agency is attributed to a system when the system's algorithms permit it to achieve efficiency in its decision-making given the limitations on the complexity required for the ways in which the system is designed to interact with its environment. From an algorithmic information theory (AIT) perspective, an agent's information base can be viewed as a compressed representation of its environment. To be successful, agents must be able to identify patterns in the environment needed for the generation of a model that is sufficiently compact to be usable. Agency in a system relies on its ability to compress the raw data to which it has access. To the extent that an agent can carry out a task while minimizing the data it uses to do so, it will be an effective agent. The more "complex" the environment, the more sophisticated the agent's internal algorithm must be to remain effective.

In their acclaimed textbook on artificial intelligence, Stuart Russell and Peter Norvig (Russell & Norvig, 2020) describe the simplest AI agents, called reflex agents or reactive agents, as those that respond to inputs without planning, coordinating, or setting goals. They define cognitive agents as those

designed to perceive their environment, reason about it, and make decisions based on objectives. A rational agent is a type of cognitive agent that acts to maximize its expected utility, where utility is measured by the objectives set by the system's designers and the chances that a task will be satisfactorily completed. An ideal rational agent would calculate how to achieve maximum utility for any contingency; however, this would require impossibly excessive computational resources. Finite AI agency, therefore, requires a trade-off between accuracy in its responses and manageable complexity. The agent needs to select which information to compress and which can be safely ignored so that it can achieve an acceptable level of utility, or successful responses, within its computational limits.

On this view, agents are classified by the complexity of their algorithms, the inputs they can act on, and the responses or action outputs they provide. A simple agent will be designed to have a low-complexity policy, such as "if $p$, do $a$." Agents become more complex when their algorithms are used to process complex inputs that elicit high-dimensional action outputs.

Another group that is involved in the attribution of agency to AI systems is that of courts, regulatory organizations, and oversight commissions. Here, the focus is on how to assign liability for harms that result from AI use and what policies should be adopted by governments for the use and development of AI. The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organization with 38 member countries that has published the first intergovernmental standards for AI. [1] To simplify the OECD definition, an AI system is a machine-based system that receives inputs from which it infers a certain kind of output. AI agents are described by the OECD (OECD, 2026) as systems that perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts. We will devote more attention to autonomy in the next section. In AI terminology, AI agents must not be confused with agentic AI. Agentic AI involves the orchestration of several AI agents. According to the OECD:

> … agentic AI refers to systems composed of multiple co-ordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods. These systems are designed to operate in more open-ended, less predictable physical or virtual environments and to function with minimal human supervision. (OECD, 2026, p. 20).

1. https://www.oecd.org/en/topics/ai-principles.html

Before any further consideration of how agency attributions are made by regulatory institutions, another concept closely related to agency needs to be introduced: autonomy.

## Autonomy and Alignment

While Kant introduced the idea of the autonomy of the will as fundamental to moral obligation (Kant, 1785/2011, Ak. 4:431), he also recognized a sense of autonomy that could be used for corporate agents, such as a university or a state (*civitas*), to the extent they enjoy institutional self-determination (Holzhey & Mudroch, 2005, p. 59). AI systems may be considered autonomous in a sense analogous to institutional autonomy. Systems may be considered autonomous to the extent that they select from alternative courses without the selection being explicitly determined through outside intervention.

In discussions of AI, autonomy is measured by the degree to which a system can operate without ongoing human intervention. This is not a Kantian autonomy of the will. AI autonomy does not imply moral agency. There is a danger that since AI systems are designed to interact with humans like persons, moral agency will be ascribed to them. When things go wrong through the use and development of AI systems, the systems will be blamed instead of those who profit from them, design them, or use them.

Systems are said to be autonomous to the extent that they are able to complete tasks without reliance on human interventions and when their actions are not fully predictable. As the algorithmic complexity of the policy of an AI agent increases, the agent's behavior becomes less "predictable" via simple observation of the environment. As the algorithmic complexity of a system increases, there is often a conflation of "predictive autonomy" – the system's ability to operate without human intervention – with "moral autonomy." However, given our attributionist framework, this is a category error. Autonomy in AI is merely the expansion of the domain of attribution of types of agency.

Unexpected behavior from autonomous AI systems that clashes with human values has given rise to what is called the "Alignment Problem." If an AI system lies or cheats, this is considered an alignment problem. There are various strategies about how to fix it. Rules can be imposed on the system from outside (top-down approaches), or steps can be taken so that the systems themselves may develop something like norms approved by developers, oversight panels, or courts (bottom-up approaches). As AI systems become more agentic (in the sense of agentic AI frameworks), the "Alignment Problem" transforms from a technical challenge of objective-function

matching into a broader, socio-technical challenge of normative integration. Alignment, therefore, is not the pursuit of a static goal but the ongoing process of calibrating a system's "normative sensitivity" to match the shifting expectations of its diverse stakeholders. In this view, an aligned AI is not one that lacks autonomy, but one whose autonomous actions remain within the legible, revisable, and accountable bounds of human institutional intent.

However, there is another alignment problem. This is not the misalignment of particular AI behaviors with human values, but the misalignment of the effects AI systems can have on the environment, the economy, politics, and other areas of social concern. Some of these concerns are being addressed in attempts to regulate the AI industry.

## Regulation

After classifying the types of AI systems, the OECD is primarily interested in the policies that governments can adopt to ensure that rights will not be violated through the improper use of these systems. The European Union has also enacted regulations for AI, and in the official statement of 2024, the purpose of regulation is described as follows:

> The purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation. [1]

A different approach has very recently been taken by the US government. Instead of legal restrictions and liabilities, tech giants and AI labs are to police themselves, as described in the "White House Accord on Super Intelligence," which was signed by Donald Trump and executives of Google, Anthropic, Meta, OpenAI, xAI, and Nvidia on September 29, 2026 [2].

China does not have a single comprehensive horizontal law like the EU AI Act. Instead, it uses administrative measures issued mainly by the Cyberspace Administration of China (CAC) and other agencies, although a comprehensive national AI law is under legislative research and could be enacted as early as

1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02024R1689-20260727
2. https://d3i6fh83elv35t.cloudfront.net/static/2026/09/accord.pdf
   Note that Trump has renamed artificial intelligence [AI] as super intelligence [SI]

2027, "to promote the healthy development and standardized application of generative artificial intelligence, safeguard national security and public interests, and protect the legitimate rights and interests of citizens, legal persons and other organizations."[1] The emphasis is on state-defined ethical norms and social order rather than individual fundamental rights in the European sense.

India lacks a single binding AI Act, instead relying on existing laws plus voluntary guidelines. The main national framework is the India AI Governance Guidelines of November 2025, which sets ethical principles, innovation pathways, and risk-mitigation measures. It explicitly seeks to avoid heavy regulation. The long-discussed Digital India Act has not yet been enacted. It emphasizes human-centric, inclusive, and responsible AI, ethical principles, and risk mitigation. Governance of AI in India is more aspirational than the EU's binding obligations or China's detailed administrative controls. Accountability and ethical considerations are highlighted, with enforcement being lighter and less centralized.

In the measures taken around the globe to advance AI and to avoid the dangers posed by what advanced systems can do, attempts are made to distribute liabilities and responsibilities to individual human beings and corporations. No matter how advanced the AI system, the system itself cannot be held accountable any more than a piece of software. Users, designers, and developers can be held responsible, as well as the corporations that produce and use AI systems. Although there is considerable debate about what guardrails should be set up by governments to avoid risks posed by AI systems, and what steps should be taken to fix the Alignment Problem, existing AI acts and regulations focus overwhelmingly on private-sector actors – providers, deployers, developers, model owners, and users. The legal instruments and regulations do not assign formal legal responsibility to governments for recklessly pursuing what has been described as a race for the development and deployment of AI systems.

## Responsibility

Sven Nyholm has addressed the issue of legal responsibility for what automated systems do (Nyholm, 2018). His primary concern is with robotic systems, automated cars, and automated weaponry. Although he does not consider more advanced forms of agentic AI, many of his insights can be extended. The harms caused by such systems have led some ethicists to claim

1. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

that there are responsibility gaps (Matthias, 2004). Responsibility gaps are alleged to occur when none of the humans involved in the use of the system that has caused harm are held to be responsible for the harm done, and yet the system itself cannot be considered responsible. [1] In addition to responsibility gaps due to the use of automated systems, there are also responsibility gaps due to corporate activities. The legal fiction of the corporation as a person may be viewed as a means of filling at least some responsibility gaps that would otherwise result from collective agency. A corporation can be blamed and subjected to legal penalties or restrictions. Since AI systems are not capable of ownership and entering into contractual relationships, they are not considered persons even in the sense of the legal fiction that corporations are.

While discussions of the "responsibility gap" in AI ethics have been dominated by the seminal work of Sven Nyholm – who rightly characterizes these gaps as situations where a system's behavior eludes traditional models of individual culpability – recent work by Herman Veluwenkamp suggests that our conceptual reliance on a singular "responsibility gap" is, in fact, structurally defective. Veluwenkamp argues that we must disaggregate the problem into two distinct, manageable categories: Epistemic Responsibility Gaps (ERG), where the complexity of the system prevents us from knowing who or what caused the action, and Control Misalignments (CM), where the structural delegation of decision-making creates an outcome that no individual actor intended. By adopting Veluwenkamp's distinction, we move away from the futile question of whether AI can be "truly" responsible, and toward a targeted inquiry: is the harm a result of an epistemic failure (which demands greater transparency), or a control failure (which demands better institutional alignment)?

## Conclusion: Toward an Attributionist Governance

The search for a singular, essentialist definition of "AI agency" has led the field into a stalemate, oscillating between anthropomorphic fears and reductive functionalism. By shifting the focus from what an AI is to how we attribute agency to it, the paradox of modern AI governance can be resolved. Agency is not a property residing in the code; it is a relational outcome – a socio-technical construct emergent from the interactions between developers, users, regulators, and the system itself.

Our analysis suggests that we must move away from searching for a "soul"

1. This is just a rough introductory characterization of the concept of a *responsibility gap*. For a critical review of several analyses and suggestions, see (Veluwenkamp, 2025).

or "will" in machine learning models and instead cultivate a robust taxonomy of attribution. When we view AI agency as an instrument for distributing responsibility – akin to organizational agency – the "Alignment Problem" loses its existential tenor and gains a practical, manageable focus. We are not aligning "agents" with our values; we are aligning our own complex, often conflicting, regulatory and moral expectations with the systems we instantiate.

The study of artificial intelligence and machine learning systems offers many rewards. It can assist us in our assessments of human agency and actions at individual and social levels. It can prompt us to reconsider issues of divine agency. One of the most valuable lessons comes from acknowledging the contingency of our own human agency. Just as the theological concept of *fāʿil bi-l-tajallī* (agency by self-disclosure) suggests that agency can be expressed through mediated channels without residing in the mediator or in supposed divine intentions, so too should we view AI as a kind of agency without its own intentions: not a divine self-disclosure, but a manifestation of human aims and institutional structures. The goal of our study should not be to discover the mystery of agency, but to ensure that attributions of responsibility function to prevent harm. Those responsible for the actions of autonomous systems include individual developers, owners, designers, and users, as well as the organizations and institutions that make AI deployment possible, not least of which are our governments.

## ▣ Conflict of Interest

The author declares no competing interests.

## References


Anscombe, G. E. M. (1959). *Intention*. Blackwell.

Aristotle. (2014). *The Nicomachean ethics* (C. D. C. Reeve, Trans.). Hackett.

Bagnoli, C. (2021). The objective stance and the boundary problem. *European Journal of Philosophy*, *29*(3), 646–663. https://doi.org/10.1111/ejop.12678

Crook, B. (2026). Risks deriving from the agential profiles of modern AI systems. In V. C. Müller, L. Dung, G. Löhr, & A. Rumana (Eds.), *The philosophy of artificial intelligence: The state of the art* (pp. 67–82). Springer.

Descartes, R. (1972). *Treatise of man*. Harvard University Press.

Ferrero, L. (Ed.). (2022). *The Routledge handbook of the philosophy of agency*. Routledge.

Holzhey, H., & Mudroch, V. (2005). *Historical dictionary of Kant and Kantianism*. Scarecrow Press.

Ibn Sina, H. (1999). *The canon of medicine*. Brigham Young University Press.

Kant, I. (2011). *Groundwork of the metaphysics of morals: A German-English edition* (M. Gregor, Trans.; J. Timmermann, Ed.). Cambridge University Press.

Kelley, M. (2026). A control theory of action. *Australasian Journal of Philosophy*, *104*(3), 689–712. https://doi.org/10.1080/00048402.2025.2521737

Legenhausen, M. (2021). Dimensions of divine freedom. *Journal of Philosophical Theological Research*, *21*(3), 73–91. https://doi.org/10.22091/JPTR.2021.7399.2618

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201-1219. https://doi.org/10.1007/s11948-017-9943-x

OECD. (2026). *The agentic AI landscape and its conceptual foundations*. OECD Publishing.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th US ed.). Pearson.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.

Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology*, *27*, Article 14. https://doi.org/10.1007/s10676-025-09823-8

Wiener, N. (2019). *Cybernetics or control and communication in the animal and the machine* (2nd reissue ed.). MIT Press.

Zhu, L., Lu, Q., Ding, M., Lee, S. U., & Wang, C. (2026). Designing meaningful human oversight in AI. *AI and Ethics*, *6* (3), Article 286. https://doi.org/10.1007/s43681-026-01147-7.