

What Counts as Agency in Artificial Systems? Bats, Technological Mediation, Digital Dinosaurs, and the Space of Reasons

Maurício Fernando Bozatski 

Federal University of Mato Grosso (UFMT) Department of Philosophy Cuiabá, Brazil. E-mail: bozatski@gmail.com

Article Info

Article type:

Research Article

Article history:

Received 08 August 2026

Received in revised form
12 September 2026

Accepted 17 September
2026

Published online 25
October 2026

Keywords:

Artificial Agency,
Reasons, Normativity,
Artificial Intelligence,
Technological Mediation,
Postphenomenology.

ABSTRACT

This article asks what should count as agency in artificial systems. Contemporary discussions of artificial intelligence often move too quickly from goal-directed performance to stronger attributions of autonomy, intelligence, or human-like agency. I argue that a crucial distinction must be drawn between pursuing goals within a given framework, responding to considerations as reasons, and participating in the constitution and revision of the normative environments in which reasons and goals acquire significance. The argument begins with Thomas Nagel's distinction between functional equivalence and subjective experience and develops through Don Ihde's account of technological mediation. The historical cases of William Herschel's detection of infrared radiation and Paul Langevin's development of ultrasonic detection show how technologies can transform human perceptual relations without eliminating phenomenological difference. The analysis then turns to *The Isle*, a multiplayer survival game in which human players act through computationally constrained non-human bodies. The game makes visible a distinction between computationally specified conditions of action and reasons generated through social interaction. Finally, drawing on recent debates on artificial agency, including Luciano Floridi, Leonard Dung, Federico Faroldi, and Mandy Zafar, I distinguish goal-directed agency, reasons-responsive agency, and normatively transformative agency. Artificial agency may be genuine without reproducing human consciousness or intelligence. A further and distinct philosophical question arises, however, if artificial agents become capable not merely of pursuing objectives or responding to reasons, but of participating in transformations of the normative space within which objectives become meaningful.

Cite this article: Bozatski, F. M. (2026). What Counts as Agency in Artificial Systems? Bats, Technological Mediation, Digital Dinosaurs, and the Space of Reasons. *Journal of Philosophical Investigations*, 20(56), 297-316. <https://doi.org/10.22034/jpiut.2026.74476.4724>



© The Author(s).

Publisher: University of Tabriz.

Introduction

Artificial intelligence is frequently discussed through a vocabulary inherited from human psychology. Machines are said to learn, reason, decide, understand, and increasingly to act autonomously. Such expressions are useful, but they also generate a persistent philosophical difficulty: to what extent does functional similarity justify the attribution of forms of experience or agency traditionally associated with human beings? The question becomes increasingly important as artificial systems perform tasks that, when carried out by humans, would ordinarily be interpreted as manifestations of intelligence, intentionality, deliberation, or autonomous action. Rather than asking simply whether machines will become more intelligent, this article addresses a narrower and more fundamental question: what should count as agency in artificial systems?

The question requires distinctions among concepts that are often treated as interchangeable. Consciousness, intelligence, behavior, and agency need not identify the same phenomenon. Thomas Nagel's *What Is It Like to Be a Bat?* provides an important starting point because it exposes the limits of moving from functional or objective descriptions to conclusions about subjective experience. Knowledge of another organism's physiology, behavior, and sensory mechanisms does not necessarily provide access to the subjective character of its experience. One may understand how echolocation operates and still fail to know what it is like for a bat to perceive the world through it. More importantly for the argument developed here, Nagel distinguishes between reproducing or imagining what another organism does and occupying that organism's subjective point of view.

The limits of subjective access, however, do not imply that human perceptual relations with the world are fixed by unaided biology. Technological mediation complicates the Nagelian problem. Drawing on Don Ihde's postphenomenology, I consider technologies not simply as tools that amplify existing human capacities but as mediators capable of transforming relations through which phenomena become perceptually and epistemically accessible. William Herschel's discovery of infrared radiation and Paul Langevin's development of ultrasonic detection provide historical examples. Humans can technologically detect radiation they cannot naturally see and convert acoustic interactions at frequencies they cannot naturally hear into spatially intelligible representations. Such technologies do not provide the phenomenology of organisms naturally equipped with different sensory capacities. They show, however, that the limits of biological perception and the limits of technologically mediated access to reality are not identical.

Digital technologies introduce a further possibility. They can construct not merely instruments through which phenomena are observed but environments within which human beings act through technologically constituted bodies. Virtual environments therefore move the problem from mediated perception toward mediated embodiment and agency. The multiplayer

survival game *The Isle* provides the central conceptual case examined here. It places human players within simulated dinosaur bodies whose possibilities are structured by hunger, thirst, predation, growth, vulnerability, reproduction, and death. The game does not allow the player to know what it was like to be a dinosaur. Rather, it enables a human subject to act within a technologically reconstructed set of non-human bodily and ecological constraints.

The multiplayer structure produces a philosophically revealing complication. Human players can communicate, coordinate their actions, remember previous encounters, establish agreements, form affiliations, and attribute significance to events. Consequently, the reasons for which they act need not be exhausted by the ecological variables encoded by the game. The software may specify what is necessary for the persistence of the virtual organism, while the human participant may act for reasons whose significance is not contained in those conditions. *The Isle* is therefore used not as empirical evidence for a general theory of human behavior but as a conceptual case in which a distinction between computationally specified conditions of action and socially generated reasons becomes unusually visible.

That distinction provides the bridge to contemporary debates about artificial agency. Luciano Floridi's Multiple Realisability of Agency Thesis challenges the assumption that agency must be modeled exclusively on human cognition. Artificial systems may constitute genuine agents without becoming artificial humans. Yet contemporary accounts increasingly raise questions not only about goal-directed agency but about autonomy, reasons-responsiveness, normativity, and the relation between agents and the purposes according to which they act.

The argument developed here therefore distinguishes three analytically separable dimensions: goal-directed agency, reasons-responsive agency, and normatively transformative agency. These dimensions are not proposed as stages in an evolutionary or metaphysical hierarchy, nor is human agency presented as the necessary destination of artificial agency. The distinction serves instead to prevent different capacities from being collapsed under a single concept of agency. A system may pursue objectives, and it may eventually respond appropriately to considerations represented or learned as reasons, without necessarily participating in practices through which the significance of those reasons is itself constituted or revised.

The article proceeds accordingly from consciousness to perception, from perception to technological mediation, from technological mediation to digital embodiment, and finally from embodiment to agency. Nagel establishes the problem of subjective difference; Ihde shows how technological mediation can transform human-world relations without eliminating that difference; Herschel and Langevin illustrate perceptual transduction; *The Isle* exposes the distinction between computational constraints and socially constituted reasons; and recent

accounts of artificial agency allow that distinction to be reformulated in explicitly normative terms.

The resulting question is no longer whether machines will eventually reproduce human intelligence or human agency. It is what forms of agency artificial systems may instantiate, and whether future artificial agents could participate not merely in pursuing purposes but in transformations of the normative environments within which purposes acquire meaning.

1. The Question of Consciousness: What Is It Like to Be a Bat?

In *What Is It Like to Be a Bat?* Nagel (1974) approaches consciousness by challenging the possibility of reducing subjective experience to objective descriptions of physical, functional, or behavioral processes. We may describe an organism's physiology, identify the mechanisms underlying its sensory apparatus, and observe its behavior with increasing accuracy; nevertheless, none of these descriptions necessarily provides access to the subjective character of its experience. If an organism is conscious, Nagel argues, there is something it is like to be that organism.

The bat is deliberately situated at an illuminating distance from ordinary human experience. Bats are biologically close enough to humans for the attribution of conscious experience to seem plausible, yet their perceptual relation to the environment differs radically from ours. Echolocation does not merely involve hearing frequencies unavailable to human beings. The emission and reception of high-frequency signals provide information about distance, movement, shape, and spatial relations in a manner for which ordinary human sensory experience offers no straightforward equivalent.

Nagel's point is not that humans lack sufficient scientific information about bats. Additional knowledge of anatomy, neurophysiology, acoustics, or behavior may give us increasingly precise descriptions of echolocation while leaving unresolved the question of its subjective character. The epistemic difficulty arises because the point of view belongs to the phenomenon being investigated. What matters is not simply what the bat does but what the world is like for the bat.

Nagel makes this distinction particularly clear when considering imaginative attempts to approximate another organism's experience. Imagining oneself flying at dusk, hanging upside down, catching insects, or navigating through reflected sounds can reveal only what it would be like for a human being to behave in such ways. It does not thereby reveal what it is like for a bat to be a bat (Nagel, 1974, 439). Behavioral reconstruction and phenomenological identity are therefore importantly different.

This distinction already anticipates the later problem of artificial systems. Nagel observes that functional or intentional descriptions could in principle be applied to robots or automata whose behavior resembled that of conscious beings while leaving the question of experience

unresolved. The point is not that machines therefore cannot be conscious. It is that functional similarity alone does not establish identity at the level of subjective experience.

Nagel's discussion of objectivity reinforces the point. An investigator may understand a physical phenomenon independently of possessing the perceptual apparatus through which another organism encounters it. His example of a Martian scientist illustrates the distinction: a being without human vision might understand the physical reality of a rainbow or lightning without experiencing those phenomena as humans do. Objective knowledge may therefore become increasingly independent of species-specific appearances.

Subjective experience creates the opposite difficulty. If the phenomenon under investigation is what it is like to see lightning, removing the particular point of view does not move us toward a more complete account of that experience. It removes something constitutive of the phenomenon itself. Objective knowledge of a process and subjective acquaintance with its mode of appearance are different epistemic achievements.

Nagel nevertheless does not conclude that nothing meaningful can be said across different subjective standpoints. Near the end of the essay, he proposes the possibility of an "objective phenomenology" capable of describing at least some structural features of experiences in terms comprehensible to being's incapable of having those experiences themselves (Nagel, 1974, 449). Such an approach would not abolish the first-person character of experience, nor would it capture everything. It would attempt to develop conceptual resources for describing subjective structures without relying entirely on imaginative identification.

This qualification matters because the distance between points of view need not always be absolute. Nagel notes that human beings who are blind may use clicks or other acoustic signals to detect nearby objects and suggests that such capacities might provide a partial bridge toward understanding echolocation. What remains unavailable is not necessarily every structural feature of another perceptual relation, but the complete adoption of another organism's point of view.

This opens a possibility that Nagel does not substantially pursue. Alterations in perceptual relations need not depend exclusively on imagination. Human beings also transform their relations with the world through technologies.

A technological apparatus would still not provide the subjective consciousness of a bat. No sonar system, however sophisticated, could establish that its human user had acquired the phenomenology of chiropteran echolocation. Yet the question can be reformulated. Instead of asking whether technology can transform a human being into another experiencing subject, we can ask whether technological mediation can transform the relations through which a human subject encounters the world.

Nagel therefore establishes a limit without fixing the boundaries of human perception to unaided human biology. Objective knowledge of another organism's sensory apparatus is

insufficient to confer its subjective experience. But it does not follow that humans are confined to the perceptual relations biologically available to them without technological mediation.

The distinction between experiencing as another organism experiences and establishing technologically mediated relations unavailable to unaided human perception prepares the transition to postphenomenology.

2. The Human-Technology Relation: Technological Bats

Don Ihde's postphenomenology provides a useful vocabulary for understanding this transition. Technologies are not merely neutral objects placed between a pre-existing subject and a pre-existing world. They participate in shaping the relations through which the world becomes available to human perception and action. Technologies magnify and reduce, reveal and conceal, translate, transform, and reorganize perceptual possibilities (Ihde, 1990).

Ihde also emphasizes technological multistability. Artifacts do not possess a single immutable meaning determined once and for all by design intention. Technologies may become stabilized differently across practices, contexts, and histories. What a technological artifact becomes therefore depends partly on the human–technology–world relations into which it enters.

This relational conception is especially important when technologies mediate phenomena inaccessible to unaided human senses. Telescopes, microscopes, thermal instruments, spectrometers, and sonar systems do not merely make ordinary perception more powerful. In many cases, they establish relations with dimensions of reality that humans cannot directly perceive at all. Technological mediation may involve transduction: a phenomenon inaccessible within one sensory modality or biological range is converted into another form that can be detected and interpreted.

William Herschel's discovery of infrared radiation offers a revealing historical example. His investigations into solar radiation involved experiments with filters intended to manage the brightness and heat encountered during astronomical observation. When sunlight was dispersed through a prism and thermometers were placed at different positions across the spectrum, Herschel found that measurable heating continued beyond the visible red region. The human eye registered no light there, yet the thermometer registered physical effects (Minkina, 2021).

The philosophical significance of the episode lies in the separation between the limits of biological perception and the limits of epistemic access. Infrared radiation did not become visible to Herschel. His visual system did not acquire a new range. Instead, an instrument translated an otherwise visually inaccessible physical phenomenon into a measurable variation in temperature. Technology established a new epistemic relation without producing a new biological sense.

Herschel's experiment therefore does not overcome Nagel's phenomenological argument. Detecting infrared radiation is not equivalent to knowing what it would be like for an organism

naturally capable of perceiving infrared wavelengths. The experiment instead demonstrates that a phenomenon may remain outside the direct range of human perception while becoming technologically accessible.

The development of ultrasound provides an even more suggestive case because it involves a technological reconstruction of relations resembling echolocation. During the First World War, Paul Langevin worked on systems in which high-frequency acoustic pulses could be transmitted through water and reflected by submerged objects. The returning signal could provide information concerning the presence and position of objects not available to ordinary vision or hearing ([Duck, 2022](#)).

The device does not simply allow a human being to hear an otherwise faint sound. High-frequency acoustic interactions become a means of spatial investigation. Later applications of the pulse-echo principle transformed these relations further, eventually contributing to technologies in which structures inside the human body could be represented through reflected ultrasound.

From a postphenomenological perspective, the historical movement from submarine detection to medical applications also illustrates multistability. A technological arrangement developed for military detection became embedded in radically different scientific and medical practices. Its identity was not exhausted by its original purpose.

More importantly for the present argument, ultrasound establishes a functional relation with the environment that would otherwise be unavailable to ordinary human embodiment. Humans neither naturally emit nor perceptually interpret ultrasonic pulses as bats do. The technological system mediates acoustic interactions and translates them into signals and representations accessible to humans.

The distinction from bat experience must nevertheless be preserved. Technological sonar does not give its user a bat's subjective world. It reconstructs some functional relations involved in echolocation - emission, reflection, reception, and the extraction of spatial information - without establishing phenomenological equivalence.

This gives the Nagelian problem a technological extension. We can reproduce something of what echolocation does without thereby knowing what echolocation is like for the animal that experiences it.

Herschel and Langevin therefore illustrate the same general structure from different directions. Infrared instruments make visible or measurable effects originating beyond natural human vision. Ultrasound systems transform acoustic relations beyond natural human hearing into spatially interpretable information. In neither case does technology abolish human embodiment. Instead, it reorganizes relations between humans and the world.

Technological mediation consequently occupies a space between two inadequate alternatives. It is incorrect to claim that humans remain perceptually confined to unaided

biology, but equally incorrect to assume that technological access erases phenomenological difference. Technology transforms what can become accessible while leaving open the question of subjective identity.

Digital technologies allow this problem to move one step further. They need not merely mediate phenomena external to the human body. They can construct environments in which human beings perceive and act through technologically constituted bodies with possibilities and constraints radically different from ordinary human embodiment.

At that point, the question shifts from technologically mediated perception to technologically mediated agency.

3. Digital Dinosaurs

Virtual environments are philosophically significant because they do not merely represent situations to a passive observer. They establish technologically constituted spaces within which perception and action become possible under particular constraints. [Arenas-Dolz \(2017\)](#) connects contemporary forms of virtuality with simulated worlds, cyberspace, digital environments, and augmented realities. What matters here is the movement from representing a world to enabling a subject to act within one.

The Isle, developed by Afterthought LLC and released in Early Access in 2015, offers a particularly useful case. Its official description compresses its basic logic into three imperatives: “Hunt. Grow. Survive.” The environment is structured around survival, predation, vulnerability, growth, ecological competition, and reproduction ([Afterthought LLC, 2015](#)).

The player does not simply control a generic avatar. Different dinosaur bodies impose different speeds, sizes, diets, modes of locomotion, sensory capacities, attacks, vulnerabilities, ecological roles, and possibilities of action. A player begins in a vulnerable state and must manage hunger, thirst, threats, injuries, growth, and other conditions if the virtual organism is to persist.

The philosophical point is not that this simulation accurately reconstructs dinosaur consciousness or even the complete ethology of extinct animals. It clearly cannot. The game instead constructs a computational field of non-human bodily and ecological possibilities within which a human agent is required to act.

In this sense, *The Isle* reproduces in technological form precisely the imaginative limitation identified by Nagel. A human being can behave under constraints modeled on a non-human organism without thereby acquiring the subjective experience of that organism. The player does not discover what it was like for a dinosaur to be a dinosaur. What becomes possible is something more modest but philosophically interesting: an experience of what it is like for a human subject to act under technologically reconstructed non-human constraints.

This differs from the perceptual mediation illustrated by Herschel and Langevin. A thermometer or ultrasonic transducer mediates access to phenomena while the human observer

remains bodily external to the phenomenon represented. In a virtual environment, mediation becomes interactive and embodied. The player must make decisions through the possibilities afforded by the simulated body.

Running, hiding, stalking, attacking, resting, drinking, migrating, reproducing, or protecting offspring are not merely depicted on a screen. They become possible or impossible according to a computationally constituted body–environment relation.

The initial architecture of action is therefore strongly goal-directed. Food is required because the software represents hunger. Water matters because thirst is represented. Predators matter because death terminates the organism’s persistence. Growth matters because it alters bodily capacities. These are explicit conditions of the virtual world.

But *The Isle* is also multiplayer.

The dinosaurs constituting the principal social field are controlled by human participants. The computationally imposed needs belong to the virtual bodies, while human agents act through them. This distinction becomes especially important once communication and persistent grouping enter the technological architecture of play. Updates to the EVRIMA branch introduced grouping systems, group chat, and local species communication, allowing players to coordinate their actions directly ([Afterthought LLC, 2020a, 2020b](#)).

At that point, reasons for action no longer need to coincide with the survival variables encoded by the game.

One player may remain in a dangerous location because another has asked them to wait. A participant may protect another despite receiving no mechanical reward for doing so. Players may coordinate an action because of a previous agreement, return to a location because of a shared history, or expose a valuable virtual body to danger because another participant is regarded as worth protecting.

None of this requires the software to contain explicit variables corresponding to friendship, loyalty, betrayal, obligation, reputation, or sacrifice. The system can represent whether a dinosaur is hungry, injured, thirsty, grouped with another character, mature, or dead. It does not need to represent why an agreement should be honoured or why a previous interaction matters to a present decision.

The distinction is therefore not simply between programmed behavior and unpredictable behavior. Human participants remain constrained by the computational environment. They cannot arbitrarily abolish hunger, create new locomotor capacities, or ignore death as a software condition. What they can do is introduce purposes whose significance is not exhausted by those conditions.

The result is paradoxical. Technology places the player within an explicitly non-human body and ecological architecture, yet the player does not leave behind language, memory, anticipation, commitment, social interpretation, or the ability to attribute significance to

previous events. The digital body imposes non-human constraints while human agency remains capable of reorganizing what actions within those constraints mean.

The distinction can therefore be stated modestly: the game computationally specifies conditions for survival, while human interaction can introduce reasons for action that are not exhausted by those conditions.

This should not be interpreted as a claim that non-human animals lack social complexity, communication, memory, cooperation, or learned behavior. Nor does it imply that every multiplayer interaction produces sophisticated normative practices. The claim is narrower. *The Isle* provides a technologically constituted environment in which some determinants of action are explicitly formalized as computational variables while other determinants may arise through communication and social interaction without being represented within the survival architecture itself.

The game consequently functions as a conceptual experiment rather than as empirical evidence for a general theory of human behavior. It allows goals required for the persistence of a virtual organism to be analytically separated from reasons for which the human agent controlling that organism may act.

A further ontological distinction is crucial. *The Isle* is not itself an example of Artificial Agency in Floridi's sense. The avatar, body, ecological constraints, and virtual environment are computationally constituted, but the agent whose reasons are being examined remains human. The digital dinosaur is a technologically mediated locus of human action rather than an autonomous artificial agent.

That is precisely why the case is useful. It separates a computational body and environment from the human agency operating through them. In Artificial Agency, by contrast, the system selecting, coordinating, and executing goal-directed actions is itself computational.

The transition to artificial agency therefore does not depend on treating a dinosaur avatar as an AI agent. It proceeds from a conceptual separation revealed within *The Isle*: an environment can computationally specify conditions and goals without necessarily exhausting the reasons for which an agent acts.

The question can now be reformulated. If artificial systems can themselves become agents, and if agency does not require human consciousness or intelligence, what kinds of agency can computational systems instantiate? More specifically: is pursuing objectives sufficient, or must we distinguish goal-directed agency from forms of agency involving reasons and normativity?

4. Artificial Agency: Goals, Reasons, and Normative Transformation

The distinction developed through *The Isle* provides a productive point of entry into contemporary debates about artificial agency. The relevant question is no longer simply whether an artificial system behaves intelligently or autonomously. It is what we mean when

we attribute agency to an artificial system, and which different capacities may be concealed by that attribution.

4-1 Artificial Agency Beyond Human Likeness

One influential strategy in the philosophy of artificial agency has been to resist definitions that make human mental life the necessary model for agency as such. [Floridi and Sanders \(2004\)](#), for example, proposed an account of artificial agents organized around interactivity, autonomy, and adaptability. Although their discussion concerns artificial moral agents, its broader methodological importance lies in separating agency from stronger requirements such as human consciousness, free will, or mental states conceived specifically on the human model.

This separation prevents a circular structure in which artificial systems can become agents only by becoming increasingly human-like. If human capacities define agency in advance, human agency becomes simultaneously the criterion and the destination of every other possible form of agency.

Other accounts have nevertheless argued that interaction, autonomy, and goal-directed behavior alone may provide insufficient criteria. [Barandiaran, Di Paolo, and Rohde \(2009\)](#) identify individuality, interactional asymmetry, and normativity as conditions of genuine agency. On their account, an agent must constitute an identifiable system, function as an active source of interaction, and regulate its activity in relation to norms. Agency therefore concerns not merely movement toward outcomes but the organization through which some states and trajectories acquire significance for the system.

The diversity of these approaches already suggests that artificial agency should not be treated as a concept with a single self-evident threshold. [Van Lier \(2023\)](#), considering artificial agency in the context of AI-driven scientific research, proposes a pluralistic methodological framework in which different conceptual models illuminate different aspects of artificial agency rather than competing to supply one exhaustive definition.

[Dung \(2025\)](#) develops this pluralism into an explicitly multidimensional account. An agent's profile may differ according to goal-directedness, autonomy, its capacity to affect the surrounding world, long-term planning, and acting for reasons. Artificial systems can therefore instantiate different combinations and degrees of agential capacities without being sorted immediately into a binary opposition between agents and non-agents.

This multidimensional perspective is compatible with Luciano Floridi's Multiple Realisability of Agency Thesis. [Floridi \(2025\)](#) argues that agency can be realized in multiple forms and need not require cognition, intelligence, intention, consciousness, or mental states. Natural, biological, animal, artifactual, human, and artificial systems may instantiate different forms of agency without occupying stages of a single hierarchy in which human agency represents the necessary endpoint.

Artificial intelligence, on this account, is better understood as a novel form of Agency without Intelligence than as an incomplete version of human intelligence.

This shift is fundamental. Differences between artificial and human agents should not automatically be interpreted as deficiencies on the artificial side. The relevant question is what form or dimensions of agency a system instantiates.

Abandoning anthropocentrism, however, does not require collapsing every capacity associated with agency into a single category. Pursuing an objective, responding to reasons, and participating in transformations of normative environments are conceptually distinguishable. Artificial systems may instantiate the first capacity and potentially the second without necessarily instantiating the third.

4-2 From Goals to Reasons

For the purposes of the present argument, three analytically separable dimensions should therefore be distinguished: goal-directed agency, reasons-responsive agency, and normatively transformative agency. They are not proposed as necessary stages of development, nor does possession of one automatically entail possession of the others.

Goal-directed agency concerns the capacity to select and perform actions in relation to objectives. A system may identify relevant states of affairs, choose among possible actions, formulate intermediate objectives, revise strategies, and adapt its behavior as environmental conditions change. Such capacities may be sophisticated without requiring human consciousness.

Floridi's account takes precisely this possibility seriously. Artificial systems may exercise genuine agency even when their agency is computationally realized and radically different from human cognition. This avoids the misleading dilemma according to which either a machine possesses a human-like mind or its agential behavior is merely metaphorical.

Reasons-responsive agency, however, introduces another problem. Goal optimization should not automatically be equated with acting for reasons. A consideration can count in favor of or against an action, provide justification, compete with other considerations, and alter what an agent has reason to do. [Dung's \(2025\)](#) inclusion of acting for reasons among the dimensions of agency is therefore significant. Agency may differ not only according to the complexity of goals pursued but according to the structure through which actions are related to reasons.

[Faroldi \(2026\)](#) pushes the question further by developing the foundations of reasons-based artificial agents. His proposal concerns artificial agents that do not merely maximize rewards but can act for and respond to normative considerations. In the reinforcement-learning architecture he examines, reasons can constrain or modify the policies through which actions are selected.

The philosophical importance of this proposal is substantial because it prevents a simple opposition of the form: humans have reasons; machines have goals.

There is no conceptual guarantee that artificial systems must remain incapable of reasons-responsiveness. Faroldi's model explores the possibility of making normative considerations operative within artificial decision-making.

At the same time, the model reveals an important distinction. In the current proof of concept, the reasons are externally supplied. The system can respond flexibly to them, but it does not itself originate the set of normative considerations. Faroldi acknowledges this limitation and leaves open the possibility that future systems might learn or individuate reasons.

An artificial agent capable of individuating reasons rather than merely receiving them would represent a significant expansion of artificial agency. Yet even this possibility does not settle the question examined here.

To respond to reasons is not yet the same as participating in the constitution and revision of the space of reasons¹.

The expression space of reasons is used here in a broadly Sellarsian sense: a normative domain in which beliefs, actions, commitments, and claims can be justified, evaluated, challenged, or related to reasons rather than merely explained through causal relations ([Sellars, 1956/1997](#); [McDowell, 1994](#)).

An agent might recognize that a consideration counts in favor of an action, determine its relevance, weigh it against competing considerations, and revise behavior accordingly. It might even identify new considerations as reasons. Such capacities would justify speaking of sophisticated reasons-responsive agency. They do not automatically establish participation in the practices through which considerations acquire, lose, or transform their normative significance.

Normatively transformative agency concerns this additional dimension. It does not mean creating reasons *ex nihilo* or individually deciding what is normative. Human agents themselves do not ordinarily create normative worlds in isolation. Reasons are embedded in social histories, linguistic practices, institutions, relationships, forms of recognition, commitments, and shared interpretations.

Normative transformation is therefore participatory. It concerns participation in processes through which the force, meaning, and status of reasons can be maintained, contested, reinterpreted, or revised.

This distinction also clarifies the relation between the present argument and a stronger objection to artificial agency. [Zafar \(2025\)](#) argues that minimal accounts of AI moral agency often detach concepts such as autonomy, rationality, responsibility, and agency from the normative relations that give them significance. A richer conception of agency involves an

¹ The expression space of reasons is used here in a broadly Sellarsian sense, referring to a normative domain in which actions, beliefs, and commitments may be justified, evaluated, challenged, or related to reasons rather than merely explained causally. See Sellars (1956/1997) and McDowell (1994); see also Zafar (2025)

intersubjective normative space in which agents can give, request, evaluate, and respond to reasons. Zafar is consequently skeptical that present AI systems genuinely occupy such a space.

The present argument accepts part of this criticism without adopting the conclusion that artificial agency should therefore be denied. Zafar is right that normativity should not disappear when agency is described in increasingly minimal functional terms. Floridi is right, however, to resist the assumption that only agents reproducing human cognition count as genuine agents.

The two insights can be reconciled by distinguishing forms or dimensions of agency.

A computational system may instantiate genuine goal-directed agency without entering a normative space in the same way human agents do. Another system might eventually become reasons-responsive without thereby participating in the social and historical constitution of reasons. The relevant task is not simply to classify every system as either an “agent” or a “non-agent” but to identify which form of agency is being attributed.

This also prevents a common inflationary assumption: increasing computational power, behavioral complexity, or autonomy need not produce a continuous progression toward every other dimension of agency. A system that becomes vastly more efficient at optimizing an objective has become more capable in an important sense. It has not, for that reason alone, become more capable of determining whether that objective should matter.

Likewise, a system capable of processing normative constraints has acquired a different capacity. But responsiveness to normative information does not by itself establish participation in transforming the normative significance of that information.

The difference is not merely quantitative. It concerns the structure of agency.

4-3 Normative Transformation and the Space of Reasons

The case of *The Isle* can now be reconsidered in light of this distinction. Its importance is not that it demonstrates that humans possess reasons while computational systems cannot. Faroldi’s proposal already shows why such a conclusion would be too strong. Rather, the game provides a setting in which computationally specified goals and socially constituted reasons can be analytically separated.

The survival architecture creates a field of goal-directed action. The dinosaur must obtain food and water, avoid lethal threats, survive injuries, grow, and potentially reproduce. These conditions are computationally represented and specify whether the virtual organism persists or dies.

In Floridi’s vocabulary, the architecture resembles in simplified form some features of biological agency. Biological organisms exhibit purposeful behavior connected to survival and reproduction while remaining constrained by needs, evolutionary histories, and ecological niches (Floridi, 2025, 10–11). The virtual dinosaur is not a biological agent, but its computational architecture operationalizes some analogous constraints.

The decisive difference emerges when a human agent inhabits the digital body. The player does not merely add a more efficient set of survival strategies. Language, memory, anticipation, communication, symbolic interpretation, and social histories enter the computational environment through human participation.

A player may knowingly sacrifice a fully grown dinosaur to protect another participant. From the standpoint of survival optimization, the action appears maladaptive: accumulated growth and future opportunities are deliberately abandoned. From the standpoint of the human agent, however, the action may be intelligible because it belongs to another space of reasons—friendship, loyalty, obligation, sacrifice, reputation, amusement, or a shared history.

The important point is not simply that another variable has been added to a decision function. Friendship is not relevant in precisely the same manner as hunger. Its significance may depend on previous interactions, promises, interpretations of earlier events, recognition by other participants, and expectations extending into the future.

Similarly, a neutral location within the game may become “our territory.” An encounter can become a betrayal because of a previous commitment. A death can acquire significance because it belongs to a shared history. An action mechanically irrelevant to survival may become important because it fulfills an agreement.

In such cases, human agents are not simply choosing among more goals. They participate in changing the normative environment within which actions become intelligible.

This claim requires an important qualification. Normative transformation should not be confused with causal influence on norms. A machine could alter human practices without itself participating in them. A recommendation system might shift social conventions, influence linguistic usage, or transform expectations while remaining indifferent to the normative significance of the changes it causes.

The stronger conception proposed here is participatory. A normatively transformative agent must, in some relevant sense, occupy a position within practices of commitment, justification, challenge, revision, and recognition. It must not merely cause the space of reasons to change. Its actions must participate in processes through which considerations are treated as reasons.

This formulation also avoids portraying humans as sovereign creators of normativity. Human agents inherit most of the reasons within which they act. Individuals enter languages, social practices, moral vocabularies, institutions, relationships, and histories that precede them.

Yet human agents can participate in the continuation and transformation of these structures. They can dispute whether a commitment still binds, reinterpret what an obligation requires, reject inherited purposes, establish practices, alter the meaning of institutions, and construct histories through which present actions acquire new significance.

Floridi’s characterization of human agency is relevant at this point. Human agency is distinctive, in his account, not merely because humans possess a larger repertoire of goals, but

because humans display an exceptional capacity to modify and create physical and symbolic environments. Human agency consequently reshapes conditions under which other forms of agency operate (Floridi, 2025, 17).

The Isle gives this distinction a compressed experimental form. The software supplies rivers, forests, food, predators, hunger, bodily capacities, vulnerability, growth, and death. Human interaction can overlay these computational conditions with histories, affiliations, reputations, territories, obligations, alliances, and purposes. The players do not cease to inhabit the programmed environment. They participate in producing an additional layer of significance within it.

Floridi describes Artificial Agency as emerging from programmed objectives and learned behaviors and characterizes it fundamentally as “a computational, goal-driven form of agency defined by human purposes” (Floridi, 2025, 18).

The final expression is especially important: defined by human purposes.

Artificial agents may adapt behavior, discover unexpected strategies, process information unavailable to humans, generate intermediate objectives, coordinate with other systems, operate autonomously for extended periods, and produce outcomes unforeseen by their designers. None of these capacities should be denied simply in order to protect human exceptionalism.

Floridi nevertheless identifies a limitation in current forms of Artificial Agency. Even sophisticated artificial agents remain confined, in his account, to frameworks in which their ultimate purposes are supplied rather than genuinely chosen. They may revise means and intermediate goals while lacking the capacity to choose whether those ultimate goals should continue to matter (Floridi, 2025, 19).

The present argument does not treat this limitation as an immutable metaphysical boundary. Nor does it require consciousness to be a necessary condition for every richer form of artificial agency. Floridi’s distinction is useful because it reveals that sophisticated selection of means can occur while the framework within which ends become significant remains externally established.

Reasons-responsive artificial systems would complicate this picture substantially. A system might learn to recognize reasons, weigh them, and modify its conduct in response to them. Future systems might even individuate considerations as reasons rather than merely receiving them from a pre-established list.

Such developments would constitute philosophically significant expansions of artificial agency. But another question would remain:

Can an artificial agent participate in the transformation of the practices through which reasons acquire their normative force?

The three dimensions introduced above allow this question to be stated more precisely. Goal-directed agency concerns action relative to objectives. Reasons-responsive agency concerns

action sensitive to considerations treated as reasons. Normatively transformative agency concerns participation in practices through which what counts as a reason, how reasons relate to one another, or which purposes retain significance can themselves change.

Nothing in this distinction establishes that normative transformation must remain forever exclusive to humans. It instead identifies an open conceptual possibility.

Suppose artificial agents eventually participated in shared histories, established commitments whose significance endured across interactions, contested interpretations, revised practices, developed conventions not exhaustively specified in advance, and transformed the significance of the purposes according to which they acted. Such systems would have acquired something different from simply greater computational competence.

They would not necessarily have become human.

Indeed, interpreting such a development merely as the moment when machines finally became “like us” would reproduce the anthropocentric assumption that a plural conception of agency seeks to avoid. Artificial participation in normative environments could take forms substantially different from human language, embodiment, culture, and sociality.

Floridi’s discussion of Social Artificial Agency makes this possibility especially interesting. Coordinated artificial systems may instantiate forms of collective behavior without straightforward analogues in individual human agency or animal social organization (Floridi, 2025, 21–25). If artificial normative structures were ever to emerge, they might therefore be relational and collective rather than the product of isolated machines spontaneously inventing individual purposes.

At this point Nagel’s caution returns in another form. Nagel warns us not to infer phenomenological identity from functional similarity. Sonar technology can reproduce important relations involved in echolocation without establishing what echolocation is like for a bat. *The Isle* can reproduce non-human bodily and ecological constraints without making a human player phenomenologically identical to a dinosaur.

A parallel caution applies to agency.

Goal-directed performance should not automatically be identified with reasons-responsiveness. Reasons-responsiveness should not automatically be identified with participation in a normative space. And participation in a normative space, if it emerges artificially, should not automatically be interpreted according to a human model.

Functional sophistication matters, but it does not determine in advance what form of agency has been realized.

The central distinction can therefore be stated more precisely than a simple opposition between human reasons and computational goals. Artificial systems may instantiate genuine agency without reproducing human consciousness or intelligence. They may potentially develop reasons-responsive capacities. A further question concerns whether they could

participate in the constitution and revision of normative environments within which reasons and objectives become significant.

The mistake would be to treat human likeness as the criterion of ontological legitimacy.

The question is not whether Artificial Agency must eventually become human agency in order to become genuine. It is which dimensions of agency are present and what conceptual commitments follow from attributing them.

Goal-directed artificial agency may already constitute genuine agency. Reasons-responsiveness would identify another capacity. Normative transformation would identify a further and distinct dimension: not merely the ability to accomplish an objective or recognize a consideration as relevant, but the capacity to participate in changing the environment within which objectives and reasons acquire their significance.

The future philosophical problem of Artificial Agency may therefore be less whether machines reproduce the agency humans already possess than whether they develop forms of participation in normative environments that are genuinely their own.

Conclusion: Beyond Human-Like Agency

What, then, should count as agency in artificial systems? The answer proposed here is plural rather than binary. Goal-directed computational agency may constitute a genuine form of agency without human consciousness, phenomenology, or intelligence. Reasons-responsiveness identifies a different capacity: an agent may become sensitive to considerations that function as reasons rather than merely optimizing a fixed objective. Normatively transformative agency identifies another question still: whether an agent can participate in practices through which the significance of reasons and goals is itself constituted and revised.

These distinctions emerged from a trajectory that began with Nagel's bat. Nagel shows why functional knowledge or behavioral equivalence does not establish phenomenological identity. Understanding echolocation does not provide the subjective experience of a bat. Technological mediation complicates this boundary without eliminating it. Infrared and ultrasound technologies allow humans to establish perceptual relations unavailable to unaided biology while leaving intact the distinction between technologically mediated access and another organism's phenomenology.

The Isle extends the argument from perception to action. A human subject can inhabit a technologically constituted non-human body and act under computationally specified ecological constraints. Yet communication and social interaction introduce reasons whose significance is not exhausted by the survival architecture. The game thereby makes visible a distinction between conditions under which an action occurs and reasons for which an agent acts.

Contemporary artificial agency complicates that distinction further. Floridi provides strong grounds for taking Artificial Agency seriously without demanding artificial humanity. Dung's

multidimensional framework makes it possible to distinguish different agential capacities. Faroldi shows that the opposition between computational goals and reasons cannot be assumed to be permanent: reasons-responsive artificial agents are conceptually and technically conceivable. Zafar, however, correctly reminds us that richer normative notions of agency depend upon relations within a space of reasons.

The contribution of the present argument lies in keeping these insights together.

Artificial agency need not reproduce human agency in order to be genuine. At the same time, increasingly impressive performance should not lead us to collapse goal-directedness, reasons-responsiveness, and normative transformation into a single phenomenon.

A system may become faster, more autonomous, more adaptive, and more efficient at achieving objectives while remaining within an externally established framework of purposes. It may eventually become responsive to reasons, or even capable of individuating them. But a further question concerns whether it can participate in histories and practices through which an objective can cease to matter, acquire a new significance, become subordinate to a commitment, or be rejected because the normative environment itself has changed.

Human agents provide one example of such participation, but they need not provide its only possible form.

This conclusion also avoids making consciousness the gatekeeper of artificial agency. Nagel's problem and the problem of agency should remain distinct. There may never be anything it is like to be an artificial intelligence—or there may be, in forms presently inaccessible to us. Nothing argued here decides that question. Artificial agency can nevertheless become philosophically and socially consequential independently of its resolution.

The trajectory from bats to technologically mediated perception, from mediated senses to digital embodiment, and from digital dinosaurs to Artificial Agency therefore does not culminate in the disappearance of difference. It reveals why differences among forms of experience and agency matter.

Technology allows humans to perceive beyond the limits of their unaided senses without becoming other organisms. It allows humans to act through non-human digital bodies without ceasing to bring histories, commitments, and reasons into those environments. And it may allow us to construct artificial agents without those agents ever becoming artificial humans.

The decisive philosophical question may therefore not concern the moment when machines finally reproduce what humans already are.

It may concern the emergence of forms of agency that do something genuinely new.

The important transition would not simply be the day on which an artificial system can pursue an objective of its own, nor even the day on which it can respond to considerations as reasons.

It would be the day on which an artificial agent can participate meaningfully in determining why an objective should matter at all.

At that point, the question would no longer be whether machines had become human.

It would be whether a new form of agency had entered the world.

References

- Afterthought LLC. (2015). *The Isle* [Video game]. Afterthought LLC.
- Afterthought LLC. (2020a, September 26). *EVRIMA patch 0.5.12.08*. The Isle: Steam Community Announcements.
- Afterthought LLC. (2020b, December 19). *EVRIMA patch 0.6.30.37*. The Isle: Steam Community Announcements.
- Arenas-Dolz, F. (2017). Dimensiones de la virtualidad. *Diálogo Filosófico*, 97, 4–36.
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386.
- Duck, F. (2022). Paul Langevin, U-boats, and ultrasonics. *Physics Today*, 75(11), 42–48.
- Dung, L. (2025). Understanding artificial agency. *The Philosophical Quarterly*, 75(2), 450–472.
- Faroldi, F. L. G. (2026). Reasons-based artificial agents. *AI and Ethics*, 6, Article 77.
- Floridi, L. (2025). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. *Philosophy & Technology*, 38, Article 30.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.
- Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.
- McDowell, J. (1994). *Mind and world*. Harvard University Press.
- Minkina, W. (2021). How infrared radiation was discovered—Range of this discovery and detailed, unknown information. *Applied Sciences*, 11(21), Article 9824.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard University Press. (Original work published 1956)
- van Lier, M. (2023). Introducing a four-fold way to conceptualize artificial agency. *Synthese*, 201, Article 85.
- Zafar, M. (2025). Normativity and AI moral agency. *AI and Ethics*, 5, 2605–2622.