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

From the Rational Requirement and Ontological Reconstitution of Human Agency to More Onerous Duties

Muhammad Heydarpour [1] **Ali Razizadeh** [2]

1. Assistant Professor, Department of Philosophy, Faculty of Literature and Humanities, University of Religions and Denominations, Qom, Iran. **(Corresponding author)** m.heydarpour@urd.ac.ir

2. Assistant Professor, Department of Media Arts, Religion and Media Faculty, IRIB University, Qom, Iran ali.razizadeh@iribu.ac.ir

Abstract

Original Research

This paper aims to show that the rational obligation to employ AI (the first thesis), combined with the transformation of the agent's practical identity (the second thesis), can predictably prepare the ground for more numerous, burdensome, or complex rational or moral duties. This is because an agent's obligations are sensitive not only to the agent's practical identity but also to the agent's epistemic and intentional state. As these dimensions of agency are expanded, the range and stringency of the agent's obligations may likewise increase. Initially, two categories of questions – normative and descriptive – are distinguished from one another. The second part of this paper pursues the thesis that if certain AI systems are trustworthy within specified domains, the human agent is not merely rationally permitted to employ AI, but in a range of significant choices and decisions, rationally required to do so. In the third part, we explain that on the one hand, artificial intelligence is not a merely neutral instrument for antecedently given human ends; on the other, human agency is neither dissolved nor reducible to AI agency. Rather, in long-term human–AI interaction, the conditions under which human beings perceive, deliberate, will, and understand themselves undergo a fundamental reconfiguration. The final section uses an example to show how a more demanding moral obligation can arise from a transformation in the agent's practical identity together with a pre-existing rational requirement.

Keywords:

onerous duties, rational requirement, ontological reconstitution, zetetic norms, practical identity.

Received: 2026/02/26; **Received in revised form:** 2026/05/23; **Accepted:** 2026/07/29; **Published online:** 2026/08/03

Heydarpour, M.; Razizadeh, A. (2026). From the Rational Requirement and Ontological Reconstitution of Human Agency to More Onerous Duties. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 233-260. https://doi.org/10.22091/jptr.2026.17879.3731

Copyright © The Author

Introduction

A wide variety of debates and topics has emerged concerning artificial intelligence (AI). Amid these discussions, one can draw a fundamental distinction between two broad categories of problems: normative and descriptive.

The first category encompasses questions such as:

1. How ought AI system designers to engineer these systems such that moral principles are not violated?
2. Can AI agents be regarded as morally responsible agents?
3. What duties do human users have in their employment of AI?
4. Do rational agents, capable of deploying AI, possess stronger reason – on balance against contrary reasons – to incorporate AI outputs as candidate considerations under scrutiny in their deliberation about significant theoretical and practical matters? If so, on what grounds?

The second category raises questions of a different sort:

1. What advances in AI can be expected, and how will they affect economic, cultural, legal, political, and military life?
2. Can AI agents possess all, or the most salient, capacities of human agents?
3. What novel possibilities, hitherto unavailable to human beings, will AI generate?
4. Which ontological conception of human–AI interaction is correct, and why?

In the second part of this paper, we intend to argue for an affirmative answer to the last question of the first category and explain its rationale. This part pursues the thesis that if certain AI systems can merit the trust of a rational agent within specified domains, the human agent is not merely rationally permitted to employ AI, but in a range of significant choices and decisions, rationally required to do so. In the third part, addressing the last question of the second category, we respond that on the one hand, artificial intelligence is not a merely neutral instrument for antecedently given human ends; on the other, human agency is neither dissolved nor reducible to AI agency. Rather, in long-term human–AI interaction, the conditions under which human beings perceive, deliberate, will, and understand themselves undergo a fundamental reconfiguration.

Normative Engagement with AI

To elucidate human normative engagement with AI, it is necessary to examine some of the pivotal mental phenomena an agent undergoes in arriving at choices. Some of what is involved in the process of normative decision-making includes: cognitive states (such as belief, knowledge, and thought), affective states (such as fear, joy, grief, and empathy), conative states (such as intention, desire, and motivation), understanding, reasoning, experience, moral intuition, virtues and vices, character, reflection or deliberation, consciousness, autonomy, and rationality.

A standard line of argument maintains that AI agents lack consciousness, autonomy, rationality, intentionality, a sense of responsibility, and the phenomenological experience of moral obligation. Given that these features are typically conceived as necessary conditions for the attribution of moral responsibility and moral agency, AI systems cannot qualify as moral agents. Conversely, another line of argument – endorsing alternative conceptualizations of the aforementioned criteria and advocating minimalist accounts of agency – seeks to ascribe a species of moral agency to AI systems. In this paper, we aim to show that even if AI agents entirely lack consciousness, autonomy, rationality, intentionality, and the like – and hence cannot be deemed moral agents – competent human agents remain rationally obligated to consult AI systems and critically consider their outputs when making significant normative life decisions.

By "significant life decisions," we refer primarily to accepting fundamental propositions as true and to the deliberation over, and selection of, worthy ends, and secondarily to the determination of means conducive to realizing these ends. By "employment" or "use," we mean the competent deployment of an AI system to generate an appropriate output that functions as a practical consideration to be factored into the process of practical deliberation. Accordingly, the AI's final output need not be treated as an all-things-considered consideration; rather, it constitutes a prima facie (or pro tanto) consideration that requires critical human validation before acquiring decisive, normative authority. (Alternatively, an agent might first arrive at an unassisted all-things-considered judgment through standard human deliberation, and subsequently invoke AI as a critical test – examining the AI's output alongside human reasoning to corroborate, disconfirm, or revise that initial judgment before settling on a definitive, endorsed stance.) Crucially, however, consulting AI systems and treating their output as a prima facie consideration cannot be regarded as an arbitrary matter or a mere subjective preference, for reasons we elaborate below.

To address this question, we may attend to the distinctive comparative advantages AI systems possess over human beings. Among the mental phenomena involved in normative decision-making, AI systems can aid us within the spheres of cognitive states, reasoning, and reflection/deliberation. Specifically, they surpass human agents in high-speed computation, processing massive datasets, categorization, analysis, interpretation, argumentation, inference, problem-solving, and operational stability. Stated differently, humans are subject to numerous cognitive and epistemic limitations that can be mitigated through AI. These limitations include cognitive biases, affective interference (e.g., anger, fear, or partiality), conative distortions, cognitive fatigue, hunger, boredom, oversight, and memory lapses. By employing AI systems, agents can access an expansive repository of human knowledge and experiential records to optimize their choices.

For the sake of clarity and to avoid disputes, we assume that the following claims are true:

1. Rationality is divided into structural rationality (internal coherence among mental states) and substantive rationality (responsiveness to reasons);
2. Rationality is further divided into theoretical rationality and practical rationality;
3. The constitutive aim of theoretical rationality is truth;
4. The constitutive aim of practical rationality is the good;
5. The primary output of theoretical rationality is a cognitive state, such as belief;
6. The primary output of practical rationality is a conative state, such as intention;
7. Theoretical rationality consists – standardly construed – in recognizing and appropriately responding to epistemic reasons;
8. Practical rationality consists – standardly construed – in recognizing and appropriately responding to practical reasons.

1) The Requirement of Theoretical Rationality

Argument 1: Zetetic Norms

Michele Palmira (2024) argues that the rational response to higher-order evidence (HOE) requires that, upon encountering such evidence, we must reopen inquiry (re-investigate) rather than immediately abandoning our current belief. When an agent possesses first-order evidence for a proposition

p, and subsequently receives higher-order evidence that sufficiently casts doubt on the reliability or epistemic credentials of her first-order evidence, she falls under a zetetic obligation to investigate whether $p$is true. This inquiry may be direct – such as re-performing the calculation – or indirect, by investigating whether the higher-order evidence is genuinely reliable. Given the conflict between treating a matter as settled and being under an obligation to re-investigate it, zetetic norms suggest that retaining the belief as settled is no longer justified (Palmira, 2024).

Palmira adduces the following case: Andre calculates the dinner bill in her head and comes to believe that the bill is $45. She then receives highly credible testimony that she may have inadvertently taken a pill that impairs mental arithmetic. This testimony constitutes higher-order evidence: evidence about her own cognitive process and the reliability of her first-order evidence. As it happens, the pill has not actually been effective. This higher-order news does not prove that the $45 verdict is false; nonetheless, it provides Andre with a compelling epistemic reason not to treat the question as settled and to reopen inquiry.

Palmira posits that rather than directly defeating the first-order belief tout court, higher-order evidence can impose a normative obligation to reopen inquiry. Because complying with this zetetic obligation precludes an agent from treating the question as settled, retaining the belief as a closed, definitive verdict becomes rationally impermissible with respect to that obligation. In this manner, Palmira explains the normative significance of higher-order evidence through the lens of norms of inquiry (zetetic norms).

A parallel argument can be reconstructed to establish the rational requirement of employing AI in an agent's significant life decisions:

- **Premise 1 (The Epistemic Principle):** Whenever a rational agent Spossesses first-order evidence for believing that p, but simultaneously has access to strong higher-order evidence indicating a high probability of error or bias in their cognitive faculties, $S$is rationally required to re-examine p.

- **Premise 2 (The Cognitive Condition of Human Agents):** In high-stakes decision-making contexts, any rational human agent S has access to strong higher-order evidence (derived from cognitive science and lived experience) demonstrating that their intuitive and computational processes are constantly susceptible to systematic bias, fatigue, prejudice, and memory blind spots.
- **Premise 3 (From 1 and 2):** In high-stakes decision-making contexts, a rational human agent S has a zetetic obligation to re-examine their

reasoning and decision-making process.

- **Premise 4 (Epistemic Consequentialism):** The obligation to re-examine is a duty that can only be satisfied by employing methods that maximize the chance of detecting existing errors and avoid replicating the initial biases.
- **Premise 5 (The Inadequacy of Exclusive Self-Reliance):** Drawing solely on the agent's own capabilities (conducted without external tools) remains prone to the same prior biases and errors; thus, it fails to optimally satisfy the condition of epistemic efficacy, necessitating the use of external and independent assessment tools or resources.
- **Premise 6 (Partial Cognitive Efficacy of AI):** AI systems, acting as a processing tool independent of the agent's biological psychology, can provide the rational agent S with overlooked options, logical inconsistencies, and counter-evidence at a low cost.
- **Premise 7 (From 4, 5, and 6):** The employment of AI systems is one of the most effective, accessible, and rational means for fulfilling the obligation of re-examination in significant human decisions.
- **Premise 8 (From 3 and 7):** Therefore, rational agents are under a rational obligation to employ AI outputs as an effective tool for re-examination and the critical assessment of evidence in the process of their belief-formation.

Argument 2: Epistemic Blame and Normativity

According to the traditional view in epistemology, epistemic norms strictly govern the response to evidence the agent already possesses; gathering further evidence or conducting research falls solely under the purview of practical, instrumental, or moral norms, rather than epistemic norms per se. However, Carolina Flores and Elise Woodard (2023) argue for the view that based on epistemic norms and obligations regarding evidence-gathering, a rational agent is required to act and inquire within their biological and informational environment in a way that harvests evidence richness, rather than merely making correct inferences based on information already present in their mind.

The traditional view assumed that (1) belief is a passive state, whereas evidence-gathering is active; (2) epistemology only concerns norms governing mental states, not actions; and (3) the obligation to gather evidence is contingent upon the agent's practical goals. Flores and Woodard demonstrate that (*1) even processing evidence at hand requires active attention, option-structuring, and volitional inference; (*2) the traditional view cannot

consistently maintain that only norms governing mental states are epistemic, as it already accepts that abandoning a belief in light of overwhelming counter-evidence is an epistemic norm governing a mental act. Thus, epistemic rationality must have a broad scope encompassing norms governing both states and acts of inquiry; (*3) just as the norms of doxastic justification are independent of an individual's goals, the norms of evidence-gathering can also be truth-oriented norms, independent of personal preferences.

Flores and Woodard's core argument is that if there is an accepted procedure for the epistemic critique of an agent's act, then we have reason to believe that there are epistemic norms governing that act. Such an accepted procedure exists; therefore, we have reason to believe that epistemic norms governing evidence-gathering exist. [1]

- **Premise 1 (Justified Epistemic Blame):** Due to inter-subjective practices of epistemic critique, if an agent, in a significant matter, refuses to consult an accessible and efficient information source (such as AI search tools), relies solely on prior evidence, and subsequently falls into error, they may intuitively and appropriately be deemed deserving of epistemic blame and susceptible to charges of epistemic rashness or negligence.
- **Premise 2 (Blame Entails Normative Reasons):** Epistemic blame is justified only if the agent has a normative reason for x(a zetetic obligation to investigate, inquire, and evaluate evidence) and has refused to act accordingly without sufficient excuse.
- **Premise 3 (Application to AI Tools):** Advanced AI systems are currently among the prominent, efficient, and accessible resources for revealing errors, overlooked aspects of a problem, and alternative hypotheses.
- **Conclusion:** Therefore, the agent is rationally obligated (has most reason) to utilize available AI tools when forming beliefs on significant matters; deliberate refusal to do so constitutes a violation of norms of epistemic responsibility and warrants rational blame.

1. In everyday interactions, our established practices of epistemic criticism and blame reveal that we characteristically withhold epistemic trust from those who fail to inquire, gather superficial evidence, judge precipitously, lazily avoid further investigation, exhibit credulity, or display confirmation bias in their search. These practices imply that for the rational agent, merely "holding a belief or reaching a decision" is insufficient; rather, one is obligated to navigate one's epistemic environment, probe, compare, and enrich one's evidence base to avoid reliance on shallow or misleading data. Consequently, in significant decision-making contexts, indifference toward robust and accessible search tools—especially those capable of revealing higher-order evidence, serious objections, or cognitive blind spots—is simply incompatible with the prevailing norms of epistemic responsibility.

Argument 3: Diachronic Obligations and Evidence-Gathering

McWilliams (2014), writing from within the evidentialist camp, argues that an agent is obligated to reflect in certain cases (where complex evidence is involved). Because this reflection itself is a method for gathering new evidence (introspective evidence, intuitions, uncovering logical relations and hidden connections between beliefs), it constitutes a form of inquiry, and thus, the agent has an epistemic obligation to inquire. Inquiry is not merely "going to a laboratory" or "field research"; rather, mental review, inference, and the rearrangement of evidence (internal reflection) are themselves important modes of inquiry and evidence-gathering. Hall and Johnson (2018) demonstrate that if our epistemic end is to believe all and only true propositions, synchronic obligations are insufficient to achieve this goal. Epistemology must also include diachronic obligations – norms that guide our actions over time toward searching for and gathering fresh evidence.

- **Premise 1 (Insufficient Evidence and Evidence-Gathering):** For every proposition $p$regarding which the agent lacks decisive justification and the possibility of error is not ruled out, the agent has a pro tanto duty to search for and gather further evidence to reduce the risk of relying on incomplete data.
- **Premise 2 (Important Matters and All-Things-Considered Requirement):** In important matters, as the cost of error increases, the threshold for legitimate termination of inquiry rises. Under such conditions, the pro tanto duty to gather evidence can turn into an "all-things-considered" requirement, which cannot be overridden by minor cognitive or temporal costs. [1]
- **Premise 3 (Transmission of Obligation from End to Means):** If an agent is rationally obligated to achieve an optimal and justified judgment/decision regarding p, and performing action A is the most effective, least-costly method, or a necessary condition, for uncovering

1. In high-stakes situations, the rational threshold for relying on available evidence rises sharply, as the cost of potential error (asymmetric error cost) is significantly heavier than the marginal cognitive costs of time spent gathering further evidence. Therefore, in such circumstances, the weight that 'cost-of-gathering' considerations hold in the balance of practical rationality diminishes considerably. As a result, the duty to gather evidence—which was a *pro tanto* obligation in normal conditions—overcomes these marginal costs when faced with high risks, becoming an *all-things-considered* requirement for the agent (see Stanley, 2005; Fantl and McGrath, 2009). This inferential transition demonstrates that the use of AI tools to enhance evidence in critical decisions is not merely an optional choice, but an obligation stemming directly from zetetic norms linked to instrumental rationality.

unknown aspects and conflicting evidence regarding p, then the agent has a rational obligation to adopt means A.

- **Premise 4 (Relative Capacity and Neutrality of AI):** AI is a low-cost system that rapidly generates and formulates overlooked hypotheses, counter-reasons, and "higher-order evidence" concerning the agent's internal cognitive biases on a scale impossible for the individual mind.
- **Conclusion:** Therefore, in significant matters and conditions of uncertainty, a rational agent has a rational obligation to correctly utilize AI as a useful tool in the process of evidence-seeking and reasons-evaluation.

2) The Requirement of Practical Rationality

Rationality is not confined to the question of "what to believe"; it extends to the norms governing "how to think" and "how to gather evidence." Given the inherent limitations on human cognitive capacities and temporal resources, we are not required to deduce every logical consequence of our beliefs or to exhaustively gather all possible evidence for a given proposition. Consequently, the appropriate standard of rationality for human beings must reflect these cognitive bounds. Practical rationality, like theoretical rationality, is subject to these human limitations. Human agents naturally have multiple goals that cannot be realized simultaneously; thus, the rational selection among competing ends is an essential component of practical agency. Failure in this domain is distinct from epistemic failure in the realm of belief. Environmental conditions and the deliberative process play a decisive role in the rationality of decision-making. Many of our rational requirements (such as the need to seek further information or to resist temptation) exist precisely because of our cognitive and volitional limitations. Were we ideal rational agents, such "compensatory rational strategies" would be unnecessary. In short, rationality in the human sphere is a dynamic, multilayered system whose "requirements" fluctuate depending on the agent's epistemic situation, environmental opportunities, deliberative history, and even psychological traits. True rationality consists in the prudent management of these very limitations (see van Roojen, 2010).

- P**remise 1 (Practical Substantive Rationality):** Contrary to reductionist views (such as instrumental rationality), practical rationality is not limited to "finding the best means for given ends"; it encompasses the rational evaluation, harmonization, and prioritization of the ends themselves. An agent's fundamental goals constitute the structure of their life, their practical identity, and the allocation of their nonrenewable resources

(time, cognitive energy, and normative commitments). Therefore, a structural or macro-level error in the formulation of final ends may undermine the overall rationality of actions that are instrumentally effective relative to those ends.

- **Premise 2 (Deficiencies of Human Practical Rationality in Goal Selection):** Rational agents encounter persistent difficulties in selecting ends based on standard ethical theories:

    1. *From a Consequentialist Perspective:* Within the consequentialist tradition, rational goal selection depends on precise calculation of outcomes. Yet human capacities for such calculations are deeply bounded. Human actions and goals not only entail long-term, indirect consequences that are impossible for the agent to assess (Lenman, 2000), but in many cases, agents lack sufficient information regarding the weights of these consequences to calculate their expected value.
    2. *From a Deontological Perspective:* Assessing whether a "maxim" of action and its implicit end can be universalized requires a kind of coherence in intentions. Furthermore, its execution necessitates internal cognitive processes – such as precise maxim-formulation, unmixed motivation, and tracking volitional consistency – requiring the agent to meticulously avoid self-deception (Audi, 1988; Sticker, 2017).
    3. *From a Virtue-Ethical Perspective:* Under stress, time constraints, and motivational conflict, human agents often fail to act according to their virtues in particular circumstances, and the cultivation of virtue itself requires a protracted process of habituation (Miller, 2003).

- P**remise 3 (Compensatory Strategies):** Whenever a rational agent becomes aware of the limitations of their cognitive faculties, normative rationality requires the adoption of "compensatory strategies." An agent who is cognizant of the severe and predictable limitations of their cognitive, affective, and volitional capacities, yet persists in relying on errors resulting from them, is guilty of "epistemic and practical recklessness."

- **Premise 4 (Relative Efficacy of AI Systems):** Advanced AI systems possess processing and analytical mechanisms that can, to some extent, mitigate the aforementioned human limitations:

    1. *Addressing Incalculability in Consequentialism:* AI possesses a superior capacity to model complex systems, forecast multibranched scenarios, and calculate consequences beyond the reach of the human agent.

2. *Addressing Self-Deception in Deontology:* Algorithmic systems lack selfish motives, hidden interests, and psychological self-deception; they can, without being affected by these weaknesses, reveal formal contradictions, unprincipled exceptions, and tensions among conflicting commitments.
3. *Addressing Reduced Moral Sensitivity in Virtue Ethics:* Since human agents struggle to act virtuously under stress, time pressure, or motivational conflict – and given that habituation requires ongoing reinforcement – AI agents can prevent agents from falling into situational traps. This is achieved through predesigned selection models or timely warnings and reminders of moral commitments. Furthermore, AI can serve as a tool for habituation by recording, tracking, and analyzing the agent's behavioral patterns over time to provide personalized feedback for moral development.

- **Conclusion:** Therefore, assuming the availability of AI systems, a rational agent is rationally obligated to incorporate AI outputs into the process of forming, formulating, and reviewing their significant goals. Refusal to employ such tools can constitute rational negligence – an unjustified reliance on human faculties, the structural limitations of which are already established.

3) Challenges and Responses

It might be argued that if an agent seeks to utilize AI, (1) their autonomy will be undermined, potentially leading to a form of psychological/epistemic dependency that erodes the capacity for independent critical judgment over the long term; (2) their approach will drift away from a humanistic perspective toward a "mechanistic" one, falling prey to algorithmic bias; (3) the multiplicity of outputs from different AI systems may cause bewilderment and confusion; (4) one cannot trust the output due to AI's high error rate, lack of genuine understanding, or the opacity of its explanatory structure; (5) it creates a requirement for constant inquiry that leads to cognitive gridlock and practical paralysis, whereas practical rationality requires limits to inquiry; (6) it necessitates the abandonment of contextual/relational knowledge and lived experience; (7) it obscures the conditions for attributing responsibility when bad outcomes occur (a blameworthy agent could excuse her conduct by pleading, "I merely followed the rational requirement to consider the output of an approved system"); and (8) it creates a dilemma: (a) if the agent must independently evaluate and verify the AI output, they must possess expertise

in the related field (rendering the reliance on AI marginal), and (b) if the agent cannot verify it but is obligated to consider it, they have effectively engaged in blind acceptance rather than rational reflection.

In response, we contend that these challenges can be effectively defused once the nature of skillful human–AI engagement is properly conceptualized:

1. **On Autonomy:** The rational agent's reflective autonomy need not be undermined, provided that sovereign executive judgment and definitive normative endorsement in both theoretical and practical domains continue to reside strictly with the human agent.
2. **On Mechanistic Bias:** Because rational agents are aware of the architectural boundaries and potential biases of machine intelligence, they can treat the AI output as an algorithmic artifact, explicitly calibrating and distinguishing it from human and intersubjective considerations.
3. **On Confusion:** As the final decision-making authority remains with the agent, they can reflect upon multiple AI outputs and, as in the past, consult with qualified human experts.
4. **On Trust:** To evaluate the reliability of AI-generated output across diverse domains, an agent may adopt a tripartite strategy: first, utilizing multiple, independent AI systems to facilitate cross-system verification; second, subjecting these outputs to critical scrutiny and empirical testing prior to final adoption, thereby determining their context-specific reliability; and third, relying upon traditional, well-established epistemological procedures for knowledge acquisition as a corrective mechanism.
5. **On Inquiry Limits:** The rational obligation to consider AI output does not imply *ad infinitum* inquiry; it remains bound by the standard requirements of practical rationality.
6. **On Contextual Knowledge:** One can utilize this low-cost, high-speed technology without disparaging or discarding contextual/relational knowledge and lived experience.
7. **On Moral Responsibility:** Since the final selection and validation remain with the agent, responsibility for the decision resides with them; they cannot evade their moral burden.
8. **On the Expertise Dilemma:** To escape the dilemma, the agent should independently verify AI output where possible. In areas where they deem the AI output worthy of consideration but lack the expertise for full

verification, they should treat it as a "suspended" consideration (a pro tanto consideration) to be weighed until further investigation or consultation with competent human experts can resolve the matter.

4) Summary of the Normative Engagement

In light of the arguments presented, one may conclude that although contemporary AI systems and their prospective successors exhibit evident limitations and design constraints, these vulnerabilities do not undermine their advantages. They can be utilized – through training and the acquisition of requisite skills – to expand the rational agent's cognitive horizon. This allows the agent to leverage a vast array of information, knowledge, and reports of human experience, thereby minimizing human cognitive biases, deficits, and errors. If an AI system merits an agent's trust within a given domain, the agent may be rationally required to appropriately deploy their skillfully generated outputs as a pro tanto or prima facie consideration in their significant theoretical and practical decisions (whether for belief formation or goal selection).

Descriptive Engagement with AI

The normative analysis established that, under appropriate epistemic and practical conditions, rational agents may be required to engage critically with AI-generated outputs rather than exclude them from significant deliberation. The descriptive question that follows is what happens when such engagement becomes sustained and integrated into human practices. The increasing incorporation of artificial intelligence into everyday activities is transforming the conditions under which individuals perceive, deliberate, act, and understand themselves. Earlier approaches to technology largely treated technological systems as instruments for realizing purposes independently formed by human agents. Contemporary AI complicates this instrumental model by generating information, interpretations, alternatives, and recommendations that enter directly into ongoing processes of human cognition and action. Phenomenological approaches have long shown that technologies can shape how human beings encounter and interpret the world rather than merely serving as neutral means (Ihde, 1990). AI intensifies this mediating function by generating new symbolic and informational possibilities within interaction itself.

Recent accounts of artificial agency and human–AI interaction therefore suggest that the relation between humans and AI cannot be adequately understood through a simple subject–object distinction. Noller (2024)

conceptualizes AI as extending human agency through an ongoing relation between human and artificial performances of intelligence, while research on human–AI collaboration indicates that agency can become distributed across processes of interaction, contribution, and attribution (Holter et al., 2026). From a broader perspective, technological mediation shapes the relations among subjects, objects, and practices rather than merely intervening between them (Verbeek, 2005). These developments become especially significant when AI moves beyond task assistance and enters the processes through which individuals formulate intentions, evaluate alternatives, construct narratives, and reflect upon themselves.

The descriptive question is therefore not whether AI replaces human agency or produces an artificial form of human selfhood. It is how sustained AI mediation reorganizes the conditions through which human beings act and experience themselves as acting and self-interpreting subjects. This section examines this transformation through four interconnected stages: it first considers the transformation of the field of human agency, then the reconfiguration of human selfhood, before moving from technological mediation toward relational constitution and finally specifying these changes as an ontological reconstitution of the human subject.

1) AI and the Transformation of the Field of Human Agency

Artificial intelligence transforms the field of human agency not merely by increasing the efficiency of existing tasks, but by entering the processes through which possibilities for action are identified, organized, evaluated, and rendered intelligible. In a conventional instrumental conception of technology, the subject forms an intention, selects a means, and employs it to realize an already determined purpose. Contemporary AI complicates this structure by generating alternatives, organizing information, identifying patterns, formulating interpretations, and anticipating possible consequences. It can therefore participate in shaping the practical possibilities from which human intentions and actions emerge. The transformation concerns not only how efficiently humans act, but also the conditions under which intentions become actionable.

This development marks a shift from technological assistance in execution toward participation in possibility formation. Individuals may approach AI with open-ended questions, incompletely formulated problems, or indeterminate objectives, after which AI can generate alternative formulations, relevant information, possible courses of action, and interpretive frameworks. Noller's (2024) account of artificial agency is relevant because it moves beyond a strict opposition between human and artificial agency and emphasizes how

artificial systems can become involved in extending and realizing human activity. Similarly, technological mediation can shape the relations through which the world becomes available for human action and interpretation rather than merely functioning as an intermediary between subject and world (Verbeek, 2005). AI intensifies this mediating function by producing new possibilities within the activity itself.

This transformation requires distinguishing decision ownership from possibility formation. A human agent may retain final decision-making authority while no longer independently determining the informational and practical field from which alternatives are drawn. AI can identify options, prioritize information, reveal patterns, and formulate possible courses of action that subsequently become objects of human deliberation. The decision remains attributable to the human agent, while the field of deliberation may be technologically structured. Holter et al. (2026) similarly emphasize that human–AI interaction can distribute agency across processes of interaction, contribution, and attribution. Agency should therefore be examined not only at the moment of choice but also through the preceding processes that determine what becomes available for perception, evaluation, and action.

The result is a technologically structured possibility space. Human agency continues to involve perception, problem formulation, deliberation, intention, action, and evaluation, but these processes increasingly unfold within environments partly organized by computational systems. AI can influence what becomes salient, which alternatives appear comparable, how problems are framed, and which consequences become relevant. This does not require attributing human-like intentionality or consciousness to AI. As Dung (2025) argues, the philosophical challenge concerns the relationship between artificial systems and the human capacities through which action and agency are understood. The central issue is therefore not whether AI has acquired human agency, but how it transforms the field within which human agency is exercised.

Within this altered field, intention itself can become relationally formed. Human purposes may be specified, revised, and rendered more determinate through iterative engagement with computationally generated possibilities. A researcher may begin with a broad question, receive alternative conceptual framings, reject some, develop others, and return to AI for further refinement. The resulting intention remains human in its experiential and practical locus, while its articulation may emerge through an interaction in which computational contributions shape the available conceptual possibilities. This can be described as intentional co-formation without implying that AI possesses intentions equivalent to those of human subjects. Floridi (2025) helps explain this process by emphasizing AI's capacity to transform informational

and semantic environments, thereby affecting not only what individuals do but also the conditions under which purposes become intelligible and actionable.

Agency consequently appears less as an isolated capacity exercised at the final moment of choice and more as a temporally distributed process. What a person does depends partly on how a problem was framed, what information became salient, which alternatives became visible, how they were interpreted, and how intentions were subsequently revised and translated into action. AI can enter several of these stages, creating an ongoing interaction rather than a single technological intervention. This does not imply a transfer of agency from humans to machines. Human and computational contributions remain asymmetrical. The human subject experiences, evaluates, commits, and acts, whereas AI generates and transforms computational possibilities. Yet these contributions can become structurally interconnected within sustained activity, producing a relationally organized form of human agency.

The transformation is therefore best understood neither as the replacement of human agency nor as the simple enhancement of pre-existing capacities. It is a reconfiguration of the temporal, informational, and relational conditions under which human agency is exercised. AI becomes significant because it can participate in forming the possibilities through which human intentions acquire determinate form. Human agency thus remains human, while the field in which it is articulated becomes increasingly shaped through sustained interaction with computational systems.

2) AI and the Reconfiguration of Human Selfhood

If the transformation of human agency concerns changing conditions of action, the reconfiguration of selfhood concerns changing conditions of self-interpretation. Selfhood is not reducible to biological identity, psychological continuity, or stable personality traits. It also involves the reflexive capacity through which individuals interpret experiences, attribute actions and commitments to themselves, and establish continuity between who they have been, who they are, and who they may become. Ricoeur's ([1990] 1992) distinction between sameness and selfhood is crucial here. Sameness concerns continuity of characteristics, whereas selfhood involves an ongoing interpretive relation through which a person recognizes themselves across temporal and experiential difference. The increasing incorporation of AI into practices of remembering, describing, interpreting, and anticipating therefore affects not only what individuals know about themselves but also the conditions through which they become intelligible to themselves.

AI's significance for selfhood can be understood through algorithmic

reflexivity. Unlike a conventional mirror, AI systems generate computationally organized representations through classification, comparison, prediction, summarization, and interpretation. They can identify recurring patterns in personal writings, summarize previous interactions, infer preferences, and formulate descriptions of behavioral tendencies. What is returned to the individual is therefore not a neutral reflection but a technologically mediated representation of the self. AI does not simply reveal an already constituted self. It can produce representations that subsequently become materials for self-interpretation, extending the broader insight that technologies can shape how subjects encounter and understand themselves (Verbeek, 2005).

This becomes particularly significant when AI-generated representations enter the narrative organization of personal experience. Human self-understanding is partly narrative because experiences acquire coherence through their relation to memories, commitments, actions, and anticipated possibilities. Ricoeur's ([1990] 1992) account of narrative identity emphasizes that selfhood develops through the interpretation of heterogeneous experiences within an intelligible temporal configuration. AI can participate in this process by organizing autobiographical materials, identifying recurring themes, summarizing personal records, and generating interpretations of patterns across previous interactions. Such representations can then be accepted, rejected, revised, or incorporated into an individual's developing self-description. AI therefore becomes an additional element within the interpretive activity through which individuals understand themselves.

This process also transforms the relation between self-recognition and self-description. An individual may experience a recurring tendency without having conceptualized it as a significant feature of the self. When AI identifies and articulates such a pattern, something previously implicit may become available for explicit reflection. The individual can then recognize the description as accurate, incomplete, or misleading and revise it accordingly. AI thus becomes part of a reflexive process through which the subject observes, evaluates, and redescribes itself. The significance lies not in attributing self-consciousness to AI, but in recognizing its increasing involvement in the conditions through which human self-consciousness is exercised. As such representations become personalized and temporally continuous, they may become embedded in ordinary practices of self-monitoring and self-interpretation.

The same dynamic extends from the interpretation of the past to the imagination of possible futures. Markus and Nurius's (1986) concept of possible selves emphasizes that individuals understand themselves partly through imagined states they might become, hope to become, or fear becoming. AI can participate in this anticipatory dimension by generating

scenarios concerning possible decisions, careers, behaviors, or forms of personal development. By reorganizing information about the past and present into representations of possible futures, AI can contribute to the imaginative space within which individuals interpret who they might become. Selfhood can therefore be understood as temporally distributed across remembered experience, present interpretation, and anticipated possibility.

Yet AI's involvement does not establish an equivalence between human and artificial subjectivity. Human beings experience, interpret, appropriate, and live the significance of these representations, whereas AI produces computational representations without becoming a first-person subject of the life represented. This distinction remains compatible with accounts of extended cognition that recognize the functional integration of external systems without collapsing the difference between human experience and technological operation (Clark & Chalmers, 1998).

The descriptive significance of AI for selfhood therefore lies not in the replacement of human subjectivity but in the reorganization of its reflexive conditions. When computational representations become recurrently integrated into remembering, self-description, interpretation, and anticipation, AI becomes structurally involved in how individuals recognize and narrate themselves across time. Human selfhood remains grounded in lived, embodied, and first-person experience, while the processes through which that selfhood is interpreted become increasingly relational and technologically mediated. AI thus contributes to a reconfiguration of selfhood by becoming part of the reflexive processes through which human beings understand not only what they have done, but who they are and who they may become.

3) From Technological Mediation to Relational Constitution

The transformation of human agency and selfhood points toward a deeper change in the human–AI relation, namely a movement from technological mediation toward relational constitution. If AI participates not only in executing human purposes but also in forming possibilities, interpretations, and self-understandings, it can no longer be adequately described as an external intermediary. Technologies have long shaped how human beings perceive and engage with the world rather than functioning as neutral instruments (Ihde, 1990). Contemporary AI intensifies this condition because generative systems can produce new informational, symbolic, and practical possibilities within the activities they mediate. The central question is therefore when technological mediation becomes sufficiently integrated into a practice that the technological relation forms part of its constitutive organization.

This distinction can be clarified by differentiating instrumental, interactive, and generative technologies. Instrumental technologies primarily serve purposes established independently of them, whereas interactive technologies respond to and modify ongoing activities. Generative AI introduces a further transformation by producing formulations, alternatives, classifications, interpretations, and projections that can influence how an activity itself is organized. Its outputs may therefore affect not only how an intention is executed but also how that intention is initially articulated. Verbeek's (2005) account of technological mediation is relevant because technologies shape the relations through which humans encounter and interpret the world. With generative AI, this mediating function becomes dynamic and recursive, as the system responds to human activity while producing new elements that alter its subsequent course.

The resulting relation is better understood as participation rather than mere mediation. Participation does not imply that AI becomes a human subject or acquires human intentionality. Rather, computational operations become dynamically integrated into an unfolding human practice. A person may formulate an objective, receive computationally generated possibilities, accept or reject them, revise the original objective, and incorporate the resulting output into subsequent action. Human and computational operations thus remain ontologically distinct while becoming functionally interconnected. This structure can be described as coupled agency, in which human purposes and computational operations become dynamically coordinated within sustained activity without becoming equivalent forms of agency. Such an account is consistent with approaches that understand human–AI activity through interaction, contribution, and distributed attribution rather than the isolated action of a single agent (Holter et al., 2026).

The significance of this coupling becomes clearer when agency and selfhood are considered together. Human actions are remembered, interpreted, attributed, and incorporated into subsequent understandings of the self. When AI participates in forming possibilities preceding action, its contributions can therefore enter later processes of self-interpretation. Computationally generated possibilities can shape action, action can become an object of self-interpretation, and revised self-understanding can influence subsequent intentions and further interactions with AI. The significance of this recursive structure does not lie in attributing subjectivity to AI, but in recognizing that computational contributions can become embedded within the temporal continuity through which human agency and selfhood develop. Ricoeur's ([1990] 1992) account of narrative identity is relevant because selfhood is constituted through the ongoing interpretation and integration of actions and experiences across time.

This recursive integration requires distinguishing influence from

constitution. Not every technological influence constitutes an ontological transformation. Human beings have always depended upon technologies, and the modification of behavior by a technological artifact does not by itself establish that the technology has become constitutive of a practice. Constitution requires stronger integration. The technology must repeatedly enter the organization of an activity, modify relations among its stages, and become part of the conditions through which the activity is performed and subsequently interpreted. The relevant distinction is therefore between episodic mediation and sustained relational integration.

This qualification is essential because a single AI recommendation or occasional interaction does not by itself reconstitute human agency or selfhood. The constitutive claim becomes descriptively plausible when AI-mediated operations acquire continuity within a practice and later stages depend upon, respond to, or presuppose earlier computational contributions. Under such conditions, AI participates in reorganizing the practical and interpretive relations through which action acquires its form and significance. This understanding is also compatible with the extended-mind tradition, which shows how external technological resources can become integrated into cognitive processes under appropriate conditions of functional organization (Clark & Chalmers, 1998).

Relational constitution therefore describes a mode of human–AI involvement in which human capacities remain distinct while their exercise becomes increasingly organized through sustained technological interaction. It does not imply a fusion of human and artificial subjects or an ontological equivalence between them. Rather, it identifies a structured interdependence between embodied human experience and computational generation. What changes is the relational architecture within which human beings act, interpret their actions, and develop self-understanding. Under conditions of sustained integration, technological mediation becomes part of the constitutive conditions through which human agency and selfhood are dynamically organized. AI therefore matters not because it replaces the human subject, but because it becomes increasingly embedded in the relations through which that subject acts and understands itself.

4) The Ontological Reconstitution of the Human Subject

The preceding analysis clarifies what is meant by the ontological reconstitution of the human subject. The claim is neither that artificial intelligence replaces the human subject nor that AI becomes a human-like subject. Rather, the transformation concerns the relational conditions under

which human beings continue to exist as acting, knowing, and self-interpreting subjects. A classical instrumental conception assumes a relatively clear distinction between the human agent and the technological object, with the subject forming an intention and technology serving as an external means for its realization. AI-mediated activity complicates this separation because computational systems increasingly participate in the formation of possibilities, organization of action, interpretation of information, and incorporation of experience into self-understanding. The human subject therefore remains human, while the conditions through which human subjectivity is enacted become increasingly relational.

This transformation should not be understood as a movement from autonomous subjectivity to technological determination. More precisely, it involves a shift from a subject conceived as primarily independently constituted toward a subject whose modes of acting and self-interpreting are increasingly relationally constituted. Relational constitution does not eliminate the distinctive mode of human existence or establish ontological symmetry between humans and AI. Human beings remain embodied, temporally situated, experientially conscious subjects whose activities unfold within a lived world. AI systems, by contrast, remain computational entities dependent upon algorithms, data, technical infrastructures, and material systems. The relational character of human–AI existence therefore preserves this ontological difference while identifying a transformation in the conditions under which distinctively human capacities are exercised.

Ontological reconstitution can consequently be understood as a transformation in the conditions through which human beings become and remain agents and self-interpreting subjects. The significance of reconstitution lies in its systemic rather than isolated character. AI increasingly enters several interconnected dimensions of human existence: it participates in the practical organization of action by generating and structuring possibilities, in the epistemic organization of experience by filtering and reorganizing information, and in the reflexive organization of selfhood by producing representations through which individuals interpret their actions, memories, characteristics, and possible futures. These dimensions are mutually connected. What a person can do affects what the person knows, what the person knows influences the interpretation of previous actions, and such interpretations contribute to subsequent intentions.

The first dimension is therefore practical reconstitution. AI increasingly participates in the formation and organization of possibilities that precede and inform action. Human action may consequently unfold within technologically structured fields in which computational systems contribute to the generation

and transformation of alternatives. The second dimension is epistemic reconstitution. AI does not merely provide information after a human has independently established a way of understanding a problem. It can also influence which information becomes salient, how it is organized, and which interpretations become available for consideration. Human knowing is thus increasingly situated within informational environments partly structured through computational operations. The third dimension is reflexive reconstitution. By generating representations of patterns, preferences, experiences, and possible futures, AI can become involved in the processes through which individuals interpret themselves and establish continuity between past experience, present identity, and anticipated possibilities.

These three dimensions of agency, knowledge, and selfhood should therefore be understood as interconnected rather than isolated. Human beings act within fields of possibility, interpret the conditions and consequences of their actions, and incorporate those interpretations into an ongoing understanding of themselves. AI can become integrated across these interconnected processes. Its significance does not derive from possessing the same capacities as the human subject, but from becoming part of the relational environment through which those capacities are exercised and connected.

The ontological reconstitution of the human subject thus describes a transformation in the architecture of human existence rather than the emergence of an artificial human subject. The human remains the embodied and self-interpreting locus of experience, while the conditions of agency, knowledge, and self-relation increasingly include technologically generated operations. What is reconstituted is therefore not human identity itself, but the relational configuration through which human beings act, understand, and interpret themselves. AI does not replace the human subject or become its ontological equivalent. It transforms the relational conditions through which the human subject exists as an acting, knowing, and self-interpreting being.

5) Summary of the Descriptive Engagement

Human intentions and decisions are formed, revised, and determined through an ongoing process of reciprocal interaction with artificial intelligence. Given AI's algorithmic character, this interaction transforms the ways in which agents narrate themselves, under the influence of structured, computational representations. Through this transformation, the subject – as a knowing, acting, and self-interpreting being – continues to exist within a fundamental relation. In other words, by technologically configuring new possibilities that precede action, structuring the informational environment, evidence, and the

framing of problems, and shaping human recognition, self-monitoring, and the formation of narrative identity over time, artificial intelligence turns the human being into a subject whose modes of action and self-interpretation are organized as "relationally constituted."

Conclusion

The rational obligation to utilize AI (the first thesis), combined with the transformation of the agent's practical identity (the second thesis), can predictably pave the way for more numerous, burdensome, or complex rational or moral duties. This is because rational or moral obligations are dependent on the agent's practical identity, as well as their epistemic and intentional status. When these states are elevated or enhanced, the agent's duties naturally become more extensive or demanding. To examine this consequence more precisely, consider the following scenario:

- **Agent (*A*):** Sarah, a major philanthropist and director of a family charitable foundation.
- **Sarah's Goal at t_1:** "Allocating a large financial endowment (tens of millions of dollars) to reduce poverty, malnutrition, and school dropout rates in a marginalized, multinational geographical region simultaneously suffering from ecological crises (drought) and social instability." (This goal possesses the properties of "high-stakes outcomes," "high probability of resource waste and intervention failure," and "incalculable complexity for the individual human mind.")

Phase One: t_1 – The Rational Obligation to Critically Consult the System

(*Realization of Thesis 1: Rational obligation arising from higher-order evidence, the mitigation of cognitive biases, and the epistemic duty of inquiry*)

- **Status at t_1:** Influenced by direct moral intuitions, emotionally charged narratives, and empathy-driven biases, Sarah is inclined to spend the entire budget on building schools and sending immediate relief packages. She perceives herself as a "traditional philanthropist with a clear moral conscience and good intentions."
- **Introduction of AI System (*S*):** System *S*(an AI capable of development modeling and massive socio-ecological data analysis, with the ability to predict resilience, local market flows, and unintended side effects over a ten-year horizon) is available.

- **Rational Obligation:**

1. *Higher-Order Evidence:* Assume that studies in behavioral economics and applied moral philosophy show that philanthropists suffer from severe cognitive biases, such as "scope neglect" and the "identifiable victim effect." Sarah possesses evidence indicating that her individual deliberative process is highly fallible in predicting systemic outcomes.
2. *Zetetic Obligation:* Given that wasting these resources means losing the opportunity to save the lives of hundreds of children, possessing "good intentions" alone does not satisfy practical rationality. Rationality requires Sarah to make sufficient use of the available data before making a significant decision.

- **Conclusion of Phase One:** Therefore, consulting S is not merely a preference or a permissible act; it is rationally mandatory. The output of Sprovides a pro tanto reason: the model predicts that building schools without securing water safety and reforming agricultural livelihood cycles will lead to rural depopulation and leave those schools largely unused within three years. Sarah critically evaluates the output and adjusts the budget according to the model.

Phase Two: The Transition from $t_1$to t_2 – Continuous Interaction and the Reconstruction of Practical Identity

(Realization of Thesis 2: Relational constitution, algorithmic reflexivity, and the transformation of self-narrative)

- **Continuous and Recursive Interaction:** Over several years, Sarah works consistently with AI analytical platforms for all development projects, humanitarian crisis forecasting, and intervention strategies:

1. *Possibility Formation:* Novel methods of philanthropy (such as conditional cash transfers or algorithmic redesign of food supply chains) are discovered and formulated in interaction and co-creation with AI's computational possibilities.
2. *Algorithmic Reflexivity:* The system shows Sarah how her past charitable preferences were influenced by a paternalistic gaze. The AI presents her with a reflective graph of her "unconscious valuation patterns."

- **Transformation of Practical Identity at t_2 (Reconfigured Practical Identity):** The description under which Sarah deems her actions meaningful undergoes a fundamental transformation:

- At t_1, her practical identity was: "A compassionate and generous benefactor who bestows her personal wealth upon the suffering based on moral sentiments and individual autonomy."
- At t_2, her practical identity has become relationally constituted: She now understands herself as: "A strategic and moral figure within a hybrid (human–machine) decision-making network, whose mission is the radical engineering of justice structures, the resilience of vulnerable communities, and the systemic optimization of scarce life-sustaining resources." Her narrative selfhood is no longer that of a "solitary savior," but rather an agency coupled with data analysis systems.

Phase Three: t_3 – The Emergence of New, More Complex, and Burdensome Rational and Moral Duties

(New rational or moral duty stemming from the fulfillment of rational obligations and a new practical identity)

- **Emerging Crisis at t_3:** Sarah's foundation intends to implement a large-scale nutrition and treatment project across diverse climates. The AI optimization models suggest an option that maximizes overall health indicators at a macro level. However, a native minority ethnic group on the region's periphery, due to physiological and structural differences, benefits the least from this proposal and is effectively marginalized.
- **The New Moral and Rational Duty at t_3:** Sarah now faces a highly complex and binding duty:

> "The duty to re-evaluate and reform AI models in favor of intergenerational justice and minority rights (algorithmic justice auditing), to resist blind computational utilitarianism, and to systematically redesign the model to guarantee the human dignity of 'invisible' groups, even if it slightly reduces the overall statistical efficiency index."

- **Why did this duty not exist before t_1 and would not have emerged without this path?**

1. *Regarding Epistemic Status:* A traditional philanthropist at t_1 remains in a state of blameless ignorance regarding these types of systemic biases and hidden distributive consequences; they lacked the capacity to process the cross-impacts of ethnic and data variables, and simply sending food fulfilled their duty. By t_3, however, due to long-term interaction with the

systems, Sarah possesses the epistemic mastery and tools to uncover "algorithmic blind spots." When epistemic capacity increases, the scope of moral responsibility broadens: she can no longer claim ignorance of the model's hidden distributive consequences.

2. *Regarding Practical Identity and Role Obligations:* Moral duties are partially rooted in the agent's practical identity. At t_1, if Sarah had failed to notice a complex side effect on a native minority, her moral identity would not have been compromised; she was merely a "single philanthropist providing direct alms." But for Sarah at $t_2$and t_3, violating this fair audit means a direct violation of her new practical identity. She has constructed her identity based on being a "regulator of structural justice and an operator of complex systems"; abandoning this duty at t_3 would cause her personality coherence and role commitments to collapse from within.
3. *Counterfactual Check:* If Sarah had not submitted to the rational obligation of consulting the AI at $t_1$and had remained within traditional charitable methods:

- Her epistemic status would not have reached the horizon of structural analysis and second- or third-order consequences;
- Her practical identity would not have been rearranged as a subject with relational constitution;
- Consequently, at t_3, no one (not even her own conscience) would have held her responsible for failing to perform an algorithmic audit of distributive justice.

This example clearly demonstrates how the rational obligation to utilize AI, and the transformation of the agent's practical identity through long-term interaction, can predictably pave the way for more numerous, burdensome, or complex rational or moral duties for the rational agent.

▣ Conflict of Interest

The authors declare no competing interests.

References

Audi, R. (1988). Self-deception, action, and will. *Erkenntnis*, *28*(2), 247–265. https://doi.org/10.1007/BF00166408

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1111/1467-8284.00096

Dung, L. (2025). Understanding artificial agency. *The Philosophical Quarterly*, *75*(2), 450–472. https://doi.org/10.1093/pq/pqae010

Fantl, J., & McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford University Press.

Flores, C., & Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, *180*(9), 2547–2571. https://doi.org/10.1007/s11098-023-01987-9

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. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5135645

Hall, J., & Johnson, A. (2018). The epistemic duty to seek more evidence. *American Philosophical Quarterly*, *55*(2), 173–186.

Holter, S., Moruzzi, C., & El-Assady, M. (2026). Towards agency in human-AI collaboration. *IEEE Computer Graphics and Applications*, *46*(1), 13–25. https://doi.org/10.1109/MCG.2025.3623892

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy & Public Affairs*, *29*(4), 342–370. https://doi.org/10.1111/j.1088-4963.2000.00342.x

Markus, H., & Nurius, P. (1986). Possible selves. *American Psychologist*, *41*(9), 954–969. https://doi.org/10.1037/0003-066X.41.9.954

McWilliams, P. (2014). Evidentialism and epistemic duties to inquire. *Pacific Philosophical Quarterly*, *95*(2), 263–282. https://doi.org/10.1111/papq.12028

Miller, C. (2003). Social psychology and virtue ethics. *Journal of Ethics*, *7*(4), 365–392. https://doi.org/10.1023/A:1026040810355

Noller, J. (2024). Extended human agency: Towards a teleological account of AI. *Humanities and Social Sciences Communications*, *11*, Article 1338. https://doi.org/10.1057/s41599-024-03849-x

Palmira, M. (2024). Higher-order evidence and the duty to double-check. *Noûs*, *58*(3), 799–824. https://doi.org/10.1111/nous.12480

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). University of Chicago Press. (Original work published 1990)

Stanley, J. (2005). *Knowledge and practical interests*. Oxford University Press.

Sticker, M. (2017). When the reflective watch-dog barks – Conscience and self-deception in Kant. *Journal of Value Inquiry*, *51*(1), 85–104. https://doi.org/10.1007/s10790-016-9554-y

van Roojen, M. (2010). Moral rationalism and rational amoralism. *Ethics*, *120*(3), 495–525. https://doi.org/10.1086/652298

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Pennsylvania State University Press.