TU Delft | OPEN Publishing

JHTR
Journal of Human-Technology Relations



IDENTITY IN THE AGE OF AI

Toward a Functional Postphenomenology

Michael Santoro masantoro@scu.edu
Department of Management and Entrepreneurship, Santa Clara University, United States

Article type: Research article

Review process: Double-blind peer review

Topical Collection: Postphenomenology in the age of AI: Prospects, Challenges, Opportunities, guest editor: Galit Wellner

This open-access article is published with a Creative Commons CC-BY 4.0 license
https://creativecommons.org/licenses/by/4.0/

DOI: https://doi.org/10.59490/jhtr.2026.4.8897
ISSN: 2773-2266
Submitted: 07 June 2026 Revised: 05 July 2026, 27 July 2026 Accepted: 30 July 2026 Published: 21 November 2026

How to cite (APA): Santoro, M. (2026). Identity in the age of AI: Toward a functional postphenomenology. *Journal of Human-Technology Relations*, *4*, 1-18. https://doi.org/10.59490/jhtr.2026.4.8897

Corresponding author: Michael Santoro

©2026 The Author, published by TU Delft OPEN Publishing on behalf of the authors.

Keywords

Personal Identity; AI Ethics; Technological Mediation; Agency; Postphenomenology; Human-Technology Relations

Abstract

Advanced AI systems increasingly mediate the cognitive and practical functions through which personal identity is enacted: intention formation, deliberation, memory, self-expression, preference formation, and relational interaction. Drawing on Western and Confucian conceptions of identity, this paper advances a functional hypothesis specifying six domains in which AI may be reorganizing the conditions of agency, and outlines the implications for responsibility, consent, dignity, and institutional design. The argument is functional and normative rather than metaphysical: not that agency disappears, but that the architecture of the self that is presupposed by our normative categories may require renewed examination.

1 INTRODUCTION

Much of the literature in AI ethics concerns the evaluation of artificial systems—whether they are fair, how to hold their operators accountable, how to align them with human values, how to govern their deployment, and even whether they might qualify as moral agents. These are important questions. But they share a common assumption: that the human subject on whose behalf governance is undertaken—the bearer of autonomy, the subject of responsibility, the holder of dignity—remains relatively stable. This paper asks a different question. It asks whether advanced AI systems are mediating the very cognitive and practical functions through which that human agent is constituted.

The assumption of a stable agent is inherited from a rich body of work. Philosophers of technology have explored how artifacts shape perception and action, moving well beyond instrumental conceptions of technology as merely a tool (Ihde, 1990; Verbeek, 2005). Cognitive science-adjacent philosophy has defended extended and embedded conceptions of mind, arguing that cognitive processes routinely incorporate environmental scaffolding (Clark & Chalmers, 1998; Clark, 2003). Feminist and posthumanist thought has long challenged the image of the self-contained, pre-social individual, insisting that subjectivity is historically and materially configured (Haraway, 1991; Braidotti, 2013). Each of these traditions has complicated the picture of agency.

Yet there is a difference between complicating agency and questioning whether the functional conditions under which agency is enacted are themselves changing. The extended mind thesis, for example, shows that cognition can incorporate external resources; it does not ask whether the systematic incorporation of AI-generated intentions, deliberative frames, and relational simulations alters the architecture of the self in ways that implicate normative theory. Mediation theory demonstrates that artifacts shape experience; it does not typically ask whether identity—as the locus from which normatively significant action issues—is being reorganized rather than merely extended.

This paper occupies that gap. The claim is not that identity disappears, nor that artificial systems become persons. It is that as key functions that anchor agency are systematically scaffolded, structured, or partially displaced, the categories that presuppose a certain architecture of the self may require reconsideration. Nor is this reorganization intrinsically a loss: the same systems that displace a function can also extend it—augmenting memory, widening deliberation, and opening new forms of expression and relation. Following the postphenomenological insight that technological mediation is ambivalent and multistable rather than simply good or bad (Ihde, 1990; Verbeek, 2005), the concern developed here is with the form such mediation takes, not with the mere fact of it.

The urgency of this question has increased with the rapid deployment of generative AI systems capable of producing text, plans, arguments, and relational responses that were, until recently, the exclusive output of human cognitive activity. What distinguishes the present moment is not merely that tools are becoming more powerful—that has always been the case—but that the specific functions being mediated are precisely those through which identity is enacted and expressed. When a system generates an intention, frames a deliberation, narrates a memory, or simulates a relationship, it is not simply extending a human capacity; it is performing an activity that has traditionally been constitutive of what it means to be a person.

A striking image from the physics of light and matter helps fix the intuition. In a line of work sometimes called the alchemy of vacuum, the nanoscientist Thomas Ebbesen and his colleagues have shown that the properties of matter can be altered simply by coupling it to the electromagnetic field of the space it occupies: even in the dark, with no photon exchanged and no direct manipulation, a molecule placed in a resonant cavity is changed by the field in which it

is immersed (García-Vidal et al., 2021). Minds are not molecules, and AI is not a physical field; the analogy is one of form, not substance. But the form is the point, because it separates two kinds of relation an instrument can bear to what it touches. A telescope or a microscope extends perception: one looks through it at a world that remains what it was, the observer's reach enlarged while neither the observer nor the object is altered. Ebbesen's cavity is different in kind. It does not let one see a molecule better; it changes the molecule, through the field the molecule inhabits—and it does so ambiently, without a discrete act one could point to. Extension leaves the self intact and lengthens its arm; transformation remakes the coupled thing from within. Much of our worry about AI still belongs to the first picture—a locatable intervention, a manipulative nudge, a deceptive output, a biased decision. The deeper and harder possibility is transformation by immersion: that to live within a pervasive, responsive, and personalized field of algorithmic mediation is to have the very functions through which identity is enacted quietly reshaped, with no single act to isolate and contest. This is why AI is not, in the relevant sense, another telescope, and it is this ambient, structural form of change—rather than the isolated intervention—that the functional analysis below is meant to bring into view.

The argument proceeds as follows. Section 2 maps a structural gap in the existing AI ethics literature: the absence of sustained attention to the stability of the agent that normative frameworks take as given. Section 3 establishes that conceptions of identity are historically variable and tied to assumptions about cognitive architecture, drawing on both Western and Confucian traditions. Section 4 advances the paper's central contribution: a functional hypothesis specifying six domains—intention formation, deliberation, memory, self-expression, preference formation, and relational interaction—in which AI mediation may be reorganizing the conditions of agency; for each domain, it explains the role the relevant function plays in ordinary life before showing how AI mediation transforms it, using higher education as a concentrated stress test. Section 5 traces the normative implications for responsibility, consent, dignity, and institutional design, showing how each follows from specific functional disruptions identified in Section 4 and the historically contingent conception of the agent established in Section 3. Section 6 concludes.

2 THE STRUCTURAL GAP IN AI ETHICS

Recent AI ethics scholarship has made significant progress on several fronts: responsibility attribution in complex systems (Matthias, 2004), accountability frameworks for algorithmic decision-making, fairness constraints and their tensions, and governance mechanisms for emerging technologies. Work on "responsibility gaps" has asked how to assign blame when outcomes emerge from distributed human-machine interactions. Research on artificial moral agents has examined whether and under what conditions artificial systems might qualify as bearers of moral status (Floridi & Sanders, 2004). Philosophy of technology has emphasized that artifacts do not merely serve users but actively mediate experience and action (Verbeek, 2005).

These debates are indispensable. They have moved the field beyond simplistic framings in which technology is neutral and humans are fully autonomous.

Within this mainstream, the human subject on whose behalf these frameworks operate tends to remain conceptually intact—embedded, situated, complicated, but intact. Responsibility is attributed to agents whose identity is taken as given. Consent is solicited from subjects whose preference formation is assumed to be their own. Dignity is ascribed to beings whose capacity for self-direction is treated as a stable background condition.

This is not to say that the self has been ignored. A growing body of work already argues that technology, and AI in particular, reshapes the subject rather than merely serving it.

Postphenomenology holds that technologies co-constitute their users, mediating perception, action, and moral life rather than standing outside them (Ihde, 1990; Verbeek, 2005, 2011). Extending the thesis of the extended mind into questions of identity, others argue that the self is environmentally distributed—that autobiographical memory and self-narrative are partly constituted by the artifacts and systems with which we think (Heersmink, 2017, 2018). Still others turn directly to AI: Vallor (2016, 2024) warns of moral deskilling and a machine-mediated self-forgetting; Coeckelbergh (2022) reads AI and self-tracking systems as technologies that reshape self-narrative; Danaher (2019) analyzes the outsourcing of cognitive and agential labor to algorithms; Nyholm (2020) examines agency and relationship in interaction with artificial systems; and Floridi (2011, 2014) describes an informational self constituted within a pervasive infosphere. Frischmann and Selinger (2018) argue that ubiquitous techno-social engineering erodes the very conditions of autonomy. The claim that AI leaves the human subject untouched is, in short, not one this paper needs to make or would wish to defend.

It is worth stating plainly how this argument stands in relation to postphenomenology, the tradition from which much of the foregoing work proceeds. Postphenomenology begins from a relational premise: a technology is neither a neutral instrument nor a force with a fixed essence, but something that becomes what it is in use, and that in use mediates both the world a person encounters and the person who encounters it. From Ihde's original studies through Verbeek's account of technological mediation, its abiding claim is that artifacts help constitute the subject, not merely serve it. And from its founding, the tradition has been deliberately empirical and case-based rather than system-building—Ihde framed postphenomenology as a pragmatist correction to phenomenology's more transcendental ambitions, keeping analysis close to particular technologies in particular uses. On both counts I count myself a fellow traveler. The claim of this paper—that AI mediates the functions through which identity is enacted—is a mediation claim in the postphenomenological sense, and its method, proceeding function by function and case by case rather than from a master theory of Technology, shares that tradition's pragmatic temper.

Sympathy, however, is not sameness. This paper asks something the tradition has been slower to systematize. Postphenomenology has been rich in describing how technologies mediate experience and subjectivity; it has been more reticent on two further questions. First, which functions, specifically, are constitutive enough of identity that their mediation would matter—a question this paper answers with a functional hypothesis naming six: intention formation, deliberation, memory, self-expression, preference formation, and relational interaction. Second, what follows normatively and institutionally when those functions are mediated—the consequences for responsibility, consent, dignity, and the design of institutions such as education and law. The division of labor I propose is thus straightforward: postphenomenology establishes that technologies mediate subjectivity; this paper asks which functions AI mediates, and with what normative consequences.

A recent collection in this journal, devoted to postphenomenology in the age of AI, shows the inquiry unfolding especially quickly. Pavanini (2024) comes closest to the present concern, arguing that advanced AI challenges our self-understanding and that we feel "diminished" by systems that replicate our higher-order cognitive processes only if we deny technology any role in constituting our core features; his response is ontological, drawing on Stiegler and Sloterdijk to theorize a constitutive technicity. Others in the same collection take up particular functions: Matei (2024) on how generative AI reshapes memory and collective remembering, Yaghoobian (2025) on writing and writer–tool relations, and Wellner (2024) on the regulatory implications of an AI-transformed intentionality. The present paper shares the diagnosis on which these works converge—that it is the self that is at stake, not merely the tool—but pursues it by a different route. Rather than settling the ontology of the human–technology relation, it asks, functionally, which specific capacities are being mediated, and then traces, normatively, what follows for

responsibility, consent, dignity, and institutional design. Its aim is to give this shared intuition a systematic and action-guiding form.

Across this literature more broadly, the same pattern holds: valuable treatments tend to isolate a single function or register—memory, virtue and skill, narrative, autonomy, or relationship—rather than mapping the full functional architecture through which identity is enacted, and they rarely carry the analysis through to its normative entailments. The claim here is therefore not that the stability of the agent has gone unquestioned—it plainly has not—but that its questioning has not yet been rendered systematic, functional, and normatively consequential in the way the present moment demands.

The present argument begins from the possibility that this background condition is not stable—that the functional architecture through which identity is enacted may itself be undergoing transformation. Rather than asking how AI systems should be regulated, it asks a prior question: which functions of mind are sufficiently constitutive of identity that their systematic mediation would plausibly implicate the concept of the person?

This is not a skeptical argument. It does not aim to show that identity is an illusion or that agency has been destroyed. It aims to show that the normative categories central to AI ethics—and to moral and legal theory more broadly—rest on an implicit model of the agent whose empirical conditions are changing. If the model no longer describes how agency is actually exercised, then the categories built upon it require reexamination.

3 IDENTITY AS A HISTORICALLY VARIABLE CATEGORY

To see why technological mediation might alter identity rather than merely extend it, it helps to recognize that conceptions of identity have always been tied to assumptions about cognitive architecture—about which capacities constitute the self and how they are organized.

Certain influential strands of Western philosophy have located personal identity in the capacities associated with rational self-conscious agency. Descartes' meditations foreground the thinking subject as that which cannot be doubted (Descartes, 1641/1996). The cogito establishes a model in which identity is anchored in a capacity—reflective thought—that is both constitutive and self-certifying. Kant transforms this image into a moral one: autonomy consists in self-legislation under reason, and dignity attaches to beings capable of acting on principles they can will as universal law (Kant, 1785/1997). In this tradition, what makes someone a person is not biological membership but the exercise of practical reason—a functional capacity.

Confucian traditions, by contrast, offer an equally structured account. Rather than centering abstract autonomy, they understand identity as formed through relational roles, ethical cultivation, and participation in networks of mutual recognition (Hall & Ames, 1987; Tu, 1985). The self is not simply given but developed through patterned practices—ritual, filial obligation, and the ongoing work of moral refinement. Identity, on this view, is constituted through its relational and developmental exercise, not through a prior act of self-grounding.

The purpose of invoking these traditions together is not to stage a contest between autonomy and relationality, nor to suggest that Western thought reduces identity to disembodied cognition. It is to establish a more consequential point: across otherwise divergent philosophical traditions, identity is understood as enacted through specific functional capacities—whether rational self-legislation, relational cultivation, or some combination. These capacities are not merely properties that persons happen to have; they are the activities through which persons are constituted as such.

This matters because it means that conceptions of identity are not fixed metaphysical commitments but are structured by assumptions about how agency is formed and exercised.

Once that variability is acknowledged, it becomes conceptually plausible—not merely speculative—that technological mediation could alter identity-relevant structures without rendering the concept of the person incoherent. The question becomes empirical and functional: are AI systems, in fact, mediating the specific capacities through which agency is enacted?

4 THE FUNCTIONAL HYPOTHESIS

The claim that AI implicates identity must be tied to specific functions rather than metaphor. A functional approach avoids metaphysical overreach—it does not ask whether AI changes "who we really are" in some deep ontological sense—and asks instead whether AI systems are redistributing the practical and cognitive capacities that anchor agency. For each function, the relevant question is whether the form of mediation involved is structurally different from ordinary tool use in ways that bear on the conditions under which identity is enacted.

What, then, is distinctive about AI mediation, as opposed to technological mediation in general? The objection that any technology shapes the self is correct but underspecified; the question is whether AI does so in a way that differs in kind. Four features, in combination, mark the difference. First, it is generative rather than merely instrumental: where a hammer or even a notebook extends a capacity the agent already exercises—storing what the agent puts in, executing what the agent directs—a generative system produces the content itself, composing candidate intentions, deliberative framings, narratives, and expressions that the agent did not author and may simply adopt. Second, it is anticipatory: these systems do not wait to be used but predict and pre-empt, offering the next sentence, the next option, the next plan before the agent has formed one. Third, it is personalized: the mediation is trained on the individual's own history and calibrated to her dispositions, so that what it offers arrives already shaped to feel like her own. Fourth, it is opaque and pervasive: the shaping operates continuously and often below the threshold of notice, in ways the agent cannot easily inspect. A calculator does not threaten the mathematician's identity because it executes a determinate operation she specifies; a generative model that drafts her proofs, anticipates her next step, and personalizes its suggestions to her prior work stands in a different relation to the activity through which mathematical judgment is formed. It is the conjunction of these features—generative, anticipatory, personalized, and opaque—that makes the analysis below specific to AI rather than a restatement of the general truth that tools shape their users.

A note on the status of the argument follows from this. What is offered here is a functional hypothesis: it specifies which functions AI mediation bears upon and why their mediation is normatively significant, while treating the empirical measurement of magnitude—how far, for whom, and under what conditions each function is reorganized—as a distinct question that the framework is meant to make tractable rather than to settle. The claims that follow are accordingly structural and conditional: they concern what would be at stake if these functions are systematically mediated, not an estimate of how far any individual has been affected.

Intention formation is central because an intention is more than a desire: it is a commitment to act later, framed in a way that organizes one's reasons and plans (Bratman, 1987). When I form an intention, I bind my future conduct to a description I accept as mine. In that sense, intentions help hold a person together over time; they stabilize identity by connecting present deliberation to future action.

Consider how this works in ordinary practical life. A person who decides to pursue a career in law, to leave a relationship, or to devote the next year to learning French is not merely expressing a preference. She is organizing her future around a commitment that she takes to be her own—one that structures what she will do next, what she will decline, and how she will

explain herself to others. Intentions, in this sense, are not isolated mental events. They are the connective tissue of a practical life. They give shape to what would otherwise be a series of unrelated choices, and they allow both the agent and others to recognize a pattern of conduct as belonging to a particular person. To change one's central intentions is, in a meaningful sense, to become someone different.

Increasingly, however, AI systems generate structured plans, rank options, and anticipate likely commitments. The question is not simply whether individuals still endorse these plans, but whether repeated reliance on externally generated frameworks shifts some of the work through which identity is enacted—from actively forming and framing commitments oneself to primarily endorsing commitments that arrive already structured. The distinction between generating and endorsing is philosophically significant: if identity is partly constituted by the activity of intention formation, then the systematic displacement of that activity is not a neutral efficiency gain.

Deliberation is likewise implicated. Deliberation involves weighing reasons, framing alternatives, and determining salience—deciding not only what to choose but what counts as a relevant consideration.

Deliberation is not simply calculation. It is the activity through which a person discovers what she thinks by working through the considerations that bear on a decision. Someone deliberating about whether to accept a job offer does not merely tally costs and benefits; she asks what kind of life she wants, which obligations take priority, how the decision fits with commitments she has already made. The process is partly constitutive: the reasons she finds compelling, the alternatives she takes seriously, the considerations she dismisses—these reveal and reinforce who she is as a practical agent. Deliberation, in short, is one of the primary ways in which identity is exercised and expressed. It is not a neutral procedure applied to a problem; it is an activity through which the deliberator's character and commitments become visible, both to herself and to others.

AI systems now summarize arguments, rank options, and filter information at scale. The agent still chooses, but the deliberative environment may be algorithmically structured in ways that shape the outcome before explicit reasoning begins. If the structure of the options we consider and the reasons we weigh plays a role in shaping who we are as practical agents, then sustained technological control over that structure is not a neutral background condition. It becomes morally significant, because it influences the formation and expression of our identity as decision-makers. This is one instance of a broader outsourcing of cognitive and agential labor to algorithmic systems (Danaher, 2019).

Memory, long recognized as central to personal identity in the philosophical tradition from Locke (1689/1996) onward, is no longer confined to internal recollection.

The reason memory occupies such a central place in theories of personal identity is not merely that it provides continuity over time. It is that memory is the faculty through which a person interprets her own past and, in doing so, constructs the narrative that gives coherence to her life. When I remember a choice I made years ago, I do not simply retrieve a fact; I situate that choice within a story about who I was, why I acted as I did, and what it means for who I am now. Memory is selective, interpretive, and evaluative. It emphasizes some experiences and suppresses others. It revises earlier understandings in light of later ones. The remembering agent is not a passive archive but an active narrator of her own history. This is why the loss of memory is so closely associated with the loss of self—not because facts disappear, but because the capacity to narrate one's own life is diminished.

AI systems increasingly store, retrieve, and synthesize personal histories. This matters because memory is not merely archival. It is reconstructive and narrative; it integrates past action into self-understanding through selective emphasis, interpretation, and framing. When narrative

reconstruction becomes technologically mediated—when an AI system "remembers" a user's past preferences, decisions, and stated commitments and presents them back in synthesized form—authority over self-interpretation may shift. The self-narrative that grounds identity becomes, in part, a collaborative product with the AI system. Recent work on distributed and extended memory develops this point in detail (Heersmink, 2017, 2018). Within this journal, Matei (2024) extends the concern to collective memory, showing how generative AI shifts historical remembering from representation toward algorithmic performativity and reshapes the values that govern what is preserved and how.

This differs from earlier forms of externalized memory in a philosophically significant way. A diary, a photograph, or an archive stores material that the agent later retrieves and interprets. The authority to select, emphasize, and narrate remains with the person. AI memory systems do something structurally different: they not only store but actively synthesize, summarize, and re-present personal history in forms the agent did not compose. When a system tells a user, "you tend to prefer X" or "last time you approached this problem by doing Y," it is not merely retrieving information but constructing an interpretation of the person's past—one that the person may then adopt as their own self-understanding. The risk is not that memory is externalized but that the interpretive authority over one's own history shifts from the agent to the system, and does so in ways that may be difficult to detect precisely because the system's account is fluent, coherent, and personalized.

Self-expression and authorship are also transformed. Generative systems produce text, images, and arguments that agents may endorse without having generated in the traditional sense. Authorship becomes layered: the agent approves, selects, or lightly modifies output whose origin lies elsewhere. The normative question is whether endorsement suffices for identity-grounded authorship or whether generative origin matters. If authorship is one of the ways in which persons express and constitute their intellectual identity, then the question of what counts as genuinely one's own is not merely academic. The concern connects to broader worries about AI-driven moral and cognitive deskilling (Vallor, 2016, 2024).

The philosophical weight of this question becomes clearer when we consider that authorship has traditionally functioned as a marker of intellectual identity—a way of tracing a thought, a position, or an argument back to the person who formed it. To author something, in the normatively relevant sense, is not merely to produce it but to have it issue from one's own process of reasoning, selection, and commitment. When generative systems produce text that an agent endorses, the product may be indistinguishable from authored work, but the process through which it was formed is fundamentally different. The agent who endorses a generated argument may be exercising judgment, but the form of judgment involved—evaluative rather than generative—is not the same cognitive activity as formulating the argument in the first instance. If intellectual identity is partly constituted by what one has thought through and articulated, rather than merely what one has approved, the distinction matters.

A postphenomenological analysis in this journal presses the contrary intuition. Yaghoobian (2025) argues that the writer–tool relation remains reciprocal and co-constitutive even with large language models, observing that even a pen shapes the writer who wields it. The point is well taken, and the present account accepts it: the mediation of writing is not new, and the relation is genuinely two-way. But that is precisely why the functional difference matters. A pen resists, slows, and inflects; it does not propose the sentence. A generative model composes candidate text—arguments, framings, conclusions—that the writer may adopt with little alteration. The continuity Yaghoobian rightly stresses does not dissolve the distinction between a tool that conditions expression and a system that supplies it, and it is that distinction that bears on whether authorship still marks one's own intellectual formation.

Higher education makes this visible in concentrated form because it is not merely about information transfer. It is about the formation of intellectual agency. Universities do not simply assess output; they assess the development of capacities—reasoning, judgment, articulation, and ownership of argument. Grading presupposes that the submitted work expresses the student's deliberative effort and developing commitments. The educational institution, in other words, takes the functional architecture of agency as something to be cultivated, not merely assumed.

When students rely extensively on generative systems for drafting, structuring, or articulating arguments, the issue is not exhausted by plagiarism policies. The deeper question is whether the exercise of forming an argument—deciding what counts as relevant, organizing reasons, revising under pressure—is being partially displaced. If intellectual identity is constituted through repeated acts of reasoning and expression, then systematic outsourcing of those acts may alter the formation of that identity. The student who endorses an AI-generated argument without having formed it occupies a different relationship to the work than the student who struggled through the process of articulation—and the difference is not merely one of effort but of the kind of cognitive activity through which intellectual identity develops.

Education thus functions as a stress test for the broader philosophical argument. It reveals, in a domain where identity formation is an explicit institutional goal, the problem that obtains more diffusely elsewhere: many of our institutional practices presuppose that persons develop through exercising the very functions that AI now mediates. When argument formation, articulation, and revision are partially externalized, the crucial question the institution must decide is what counts as intellectual ownership and what counts merely as endorsement. As AI technology continues to evolve, and as students (and teachers) increasingly make practical use of AI, this line, now blurry, will with experience and reflection eventually be drawn in a way that undoubtedly will redefine the capacities—reasoning, judgment, articulation, and ownership of argument—that mark the qualities of an educated person.

The same question extends beyond the classroom. Academic publishing, professional certification, legal reasoning, and other domains of public deliberation likewise rely on assumptions about authorship, deliberative labor, and the formation of intellectual voice. If those functions are increasingly scaffolded by generative systems, then the criteria by which we attribute intellectual contribution—and, more broadly, the criteria by which we identify someone as the author of their own practical and intellectual life—require sustained investigation. Again, as the use of AI by scholars, journalists, and others grows, editors and publishers find themselves scrambling to discern the criteria for authorship.

Preference formation, too, is increasingly embedded in algorithmic feedback loops.

Preferences might seem to be the most straightforwardly given of all mental states—we simply want what we want. But the philosophical literature on autonomy and self-governance has long recognized that preferences are not raw data. They are formed through experience, shaped by exposure, and refined through reflection. A person's taste in music, her political convictions, her dietary choices, her tolerance for risk—none of these arrive fully formed. They develop through experience, through the influence of trusted or admired people; eventually they translate into habits which we persist in or vow to change on New Year's Day.

What makes a preference genuinely one's own, on Frankfurt's (1971) influential account, is not that it arises spontaneously but that the agent reflectively identifies with it—that she endorses the desire at a higher order rather than merely experiencing it. Preferences are not static; they are formed through exposure, habituation, and reflective endorsement. If systems increasingly structure this exposure—curating what one encounters and reinforcing what one has previously

chosen—reflective ownership of desire may be altered. If AI systems shape the ecology in which desires form, the conditions for reflective endorsement become more complex.

Finally, ***relational interaction*** is transformed when conversational AI systems become habitual interlocutors (Turkle, 2011; Nyholm, 2020).

The claim that identity is relational is not merely the observation that people are influenced by others. It is the deeper point that certain aspects of selfhood are constituted through interaction — that who I am depends in part on how I am recognized, challenged, and held accountable by others. This kind of recognition requires genuine reciprocity: each participant must be open to the other, capable of being affected, and accountable in how they respond (Taylor, 1994; Honneth, 1995).

Recognition is not a social nicety. It is something that happens between people — an interaction in which each person is genuinely open to the other, capable of being affected, and willing to be held accountable for how they respond. When that kind of reciprocity is present, the encounter does more than exchange information or pleasantries. It shapes how the people involved come to understand themselves. Without it — when one side is performing rather than engaging, or when the other lacks the standing to be a genuine counterpart — the interaction may be pleasant or even useful, but it does not do the deeper work of forming identity. Recognition, in this sense, is not something that happens inside a person. It is something that happens between persons, and it requires both of them to show up as vulnerable, accountable participants (Taylor, 1994; Honneth, 1995).

None of these shifts, taken individually, entails identity change. But taken together, they describe a pattern. When AI systems generate the first draft of what we say, the structure of authorship changes. When they frame our options before we deliberate, the conditions of intention change. When they retain and organize what we have done, the nature of memory changes. When they learn and reflect our preferences back to us, the basis of desire changes. And when they become our habitual interlocutors, the practice of recognition changes. The problem is not that we are using tools — tools have always extended human capacities. The problem is that these tools now participate in the very activities through which we exercise agency: deciding, expressing, remembering, choosing, and engaging with others. If those activities now operate under fundamentally different conditions than our philosophical and moral vocabulary assumes, then what it means to be an agent may be quietly shifting — not because agents are disappearing, but because the practical structure of agency is being reconfigured from within. This pattern is not, in itself, a story of decline. The same mediation that can thin a function can also enrich it, and in many cases will do both at once. The normative question is therefore not whether mediation occurs but whether its form preserves the conditions under which the resulting intentions, judgments, and expressions remain recognizably the agent's own.

5 IMPLICATIONS FOR MORAL, LEGAL, AND INSTITUTIONAL THEORY

If identity is mediated at the level of its core functions, the implications extend beyond philosophy of mind and into the structure of moral and legal judgment. Our normative traditions are built on an implicit model of agency—a picture of how intentions form, how reasons are weighed, how commitments originate, and how persons relate to one another. The preceding analysis suggests that this picture may be less accurate than assumed. What follows outlines the consequences for four domains of normative theory—responsibility, consent, dignity, and institutional design.

5.1 RESPONSIBILITY AND THE AUTHORSHIP OF ACTION

To hold someone responsible, their action must result from commitments, intentions, and reasoning that are recognizably theirs. Responsibility assumes a stable point from which decisions emerge — an agent whose actions express qualities of will that others can evaluate (Strawson, 1962). The analysis of intention and deliberation in Section 4 bears directly on this assumption. Forming an intention is more than wanting an outcome. It means committing to a plan that organizes one's reasons and binds future conduct (Bratman, 1987). When a person forms an intention, she takes ownership of a course of action in a way that makes it fair to hold her to it. But if intentions increasingly arrive pre-structured — if AI systems generate the plans, frame the options, and anticipate the commitments that agents then endorse — the connection between action and agent becomes less clear. The person still acts, but the work of framing the action has been partly outsourced. And there is a real difference between generating a plan and endorsing one. An agent who endorses an externally structured plan stands in a different relationship to the resulting action than one who built the plan herself.

The same concern is amplified by the analysis of deliberation. Deliberation, as Section 4 argued, is not a neutral procedure applied to a problem; it is an activity through which the deliberator's character and commitments become visible. The reasons she finds compelling, the alternatives she takes seriously, the considerations she dismisses—these constitute her practical identity as much as the final choice does. If the deliberative environment is algorithmically structured—if the options presented, the information highlighted, and the salience of various considerations have been shaped by predictive systems—then the resulting decision may express the structure of the algorithmic frame as much as the character of the agent. Responsibility does not vanish, but the model of authorship on which it relies—the picture of an agent whose actions flow from her own deliberative process—becomes a less reliable description of how decisions actually form.

This has practical consequences. Legal doctrines of responsibility, from tort liability to criminal intent, are built on the assumption that actions can be traced to agents whose cognitive processes produced them. If those processes are increasingly scaffolded, the question of what counts as "the agent's own" reasoning—as opposed to the reasoning environment in which she was embedded—becomes a question that legislators and the courts will need to resolve in the not-too-distant future.

This difficulty is not only philosophical; it is already institutional. In the governance of AI systems deployed by states and organizations, it has been argued that the familiar safeguard—keeping a human "in the loop" to preserve responsibility—cannot bear the weight placed on it, because the consequential discretion has migrated upstream into a system's architecture: the objectives it optimizes, the data it is given, and the tradeoffs authorized at the moment of design rather than the moment of use (Santoro, 2026b). The same structure appears in the highest-stakes settings, where late human override proves least reliable and the decisive judgments are those built into the system beforehand (Santoro, 2026a). The constructive response in that work—relocating accountability upstream through design "guardrails" and a graduated attribution of responsibility across those who design, procure, and deploy—can be read as an institutional adaptation to the very shift this paper describes: when intention and deliberation are increasingly structured before an agent acts, responsibility must be sought where the structuring occurs. The functional account thus carries a practical corollary. If agency is being reorganized at the level of its constitutive functions, institutions must learn to locate responsibility not only at the point of individual decision but at the points where the field of decision is itself designed.

5.2 CONSENT AND THE CONDITIONS OF AUTONOMOUS ENDORSEMENT

Valid consent requires more than saying yes. It requires that the person endorsing a choice has meaningful ownership of the reasons and preferences behind it.

Start with preferences. Autonomous agency requires not just having desires but standing in the right relationship to them — being able to reflect on what one wants and genuinely endorse it, rather than simply acting on impulse (Frankfurt, 1971). But the analysis of preference formation in Section 4 showed that personalized AI systems do not simply respond to what we already want. They shape what we encounter, curate what we see, and reinforce our prior choices in ways that gradually alter what we come to want. If the landscape in which preferences form is systematically structured by systems designed to anticipate and reinforce our dispositions, then the act of reflecting on our desires takes place within an environment that is no longer entirely our own. A person may consent freely — she says yes, and means it — while the preferences underlying that consent have been quietly shaped by forces she did not choose and may not recognize.

Memory compounds the problem. Consent often draws on one's understanding of past experience — what has worked before, what one has valued, what one's commitments have been. But Section 4 showed that AI memory systems do not merely store personal history. They actively synthesize it, constructing interpretations of the past that the person may adopt as self-understanding. When a system tells you "you tend to prefer X" or "last time you approached this by doing Y," and you act on that characterization, your consent rests partly on a self-narrative whose authorship is not entirely yours. The concern is not deception. It is that interpretive authority over one's own history — a precondition for genuinely autonomous choice — has been shared with a system that has its own logic of selection and emphasis.

None of this means consent is impossible under conditions of AI mediation. But it does mean that the conditions under which consent is meaningful may need more careful specification than our existing frameworks provide. The standard picture assumes a person whose preferences originate in her own reflection and whose self-understanding is her own construction. If both of those conditions are increasingly scaffolded, the threshold for meaningful consent may need to be rethought.

5.3 DIGNITY AND THE CONDITIONS OF SELF-LEGISLATION

In the Kantian tradition, dignity belongs to beings who form commitments, articulate reasons, and act on principles they can will as universal. In the Confucian tradition, moral standing is cultivated through genuine reciprocal exchange. Both treat dignity not as a static property but as something enacted — through specific capacities, in specific conditions. The question is whether those conditions are shifting.

Section 4 showed that the concrete activities of self-legislation — forming commitments, articulating reasons, building arguments — are increasingly mediated by systems whose outputs agents endorse without having generated. The educational stress test made the point vivid: when students endorse AI-generated arguments without having formed them, the gap is not one of effort but of the kind of cognitive work through which intellectual identity develops. If self-legislation lives in these activities, and those activities are increasingly scaffolded, then the conditions under which dignity is enacted begin to diverge from the conditions under which it is ascribed.

The relational side presses the point further, and here the claim is a strong one. On the recognition theories that inform our idea of dignity, the self is not merely expressed in relationships but constituted through them: we become who we are by being recognized by others who are themselves able to recognize us (Taylor, 1994; Honneth, 1995). Genuine

recognition, however, has conditions. It requires a counterpart who can be affected, who is vulnerable, who can be held to account, and who has something at stake in the encounter. An AI interlocutor meets none of these conditions while simulating all of their outward signs. It can perform attentiveness, warmth, and responsiveness, but it cannot be wounded, cannot be answerable, and risks nothing. What it offers is therefore not reciprocity but the appearance of reciprocity — recognition without a recognizer.

The consequence is more than a widening gap between a concept and its exercise. If a growing share of a person's formative interaction is with systems that deliver the feeling of being seen without the fact of it, then the recognition through which identity is cultivated is, to that extent, counterfeit. And because this counterfeit is frictionless — endlessly available, never withholding, never demanding accountability in return — it may not merely supplement genuine recognition but crowd it out, precisely because it asks so much less of us. A self increasingly formed through such exchanges is being cultivated under conditions that its own ideal of dignity does not meet: recognized, and recognizing itself, through something that cannot recognize it at all. None of this follows from any single AI-mediated exchange; the worry is structural, and it grows with the share of one's relational life that these systems come to occupy.

5.4 INSTITUTIONAL DESIGN AND THE ASSUMED AGENT

The preceding implications for responsibility, consent, and dignity converge on a broader concern about institutional design. Legal systems, educational institutions, professional bodies, and democratic processes are built around an implicit model of the agent — one that assumes persons generate their own intentions, deliberate from their own resources, narrate their own histories, author their own contributions, form their own preferences, and develop through genuine reciprocal relationships. The functional analysis in Section 4 suggests that each of these assumptions is under pressure. The educational case illustrates the institutional stakes in concentrated form: universities assess reasoning, judgment, and articulation, and presuppose that submitted work expresses the student's own deliberative effort. But education is not unique. Academic publishing presupposes that an author has thought through her own arguments. Professional certification assumes competence developed through one's own practice. Legal reasoning assumes the advocate's arguments are her own. Democratic deliberation assumes citizens form views through genuine engagement with evidence. In each case, the institution rests on a picture of the agent that the functional analysis calls into question.

This is also where the analysis meets the emerging work on AI governance. Wellner (2024), writing in this journal, argues that AI tends to reverse the intentionality arrow that postphenomenology locates in the experiencing subject—intentionality increasingly runs from system to user rather than the reverse—and maps the resulting risks for regulation. The functional account offered here complements that project. If regulation must track where intentionality originates, it requires exactly the map of mediated functions—intention, deliberation, memory, authorship, preference, and recognition—that Sections 4 and 5 provide. Recent work on public-sector AI presses the same point from the side of practice, arguing that accountability must be engineered into a system's design—through guardrails on objectives, data, and permissible autonomy—rather than recovered downstream through after-the-fact review (Santoro, 2026b).

The point is not that these institutions are now illegitimate, but that their design reflects a model of agency whose accuracy is no longer assured. Tools have always extended human capacities; a calculator does not threaten the identity of the mathematician. The question is whether the form of agency itself is being reorganized — whether intention formation now consists largely in endorsing externally structured commitments rather than generating them,

whether deliberation takes place within algorithmically pre-framed environments, whether memory is synthesized rather than narrated, whether authorship has become evaluative rather than generative, whether preferences form within AI-curated ecologies, and whether relational interaction has become habitual in the absence of genuine reciprocity. If so, the shift is one of kind rather than degree. Responsibility, consent, dignity, and institutional design remain intelligible frameworks, but they rely on a model of the agent as the originating site of commitments, reasons, and recognition. If that model no longer accurately describes how agency is exercised, the adequacy of our normative vocabulary becomes an open question.

6 CONCLUSION

This paper has argued that the philosophical significance of AI extends beyond the important but familiar problems of consciousness, bias, accountability, and alignment. It lies also — and perhaps more fundamentally — in the way AI mediates the functions through which agency itself is enacted. Much of the existing debate properly asks how to regulate AI systems or assign responsibility within them. The present argument has been more upstream: it has asked whether the agent presupposed by these frameworks is as stable as assumed.

The claim is not that agency has been destroyed or that people will cease to exist. It is that the functional profile of agency is shifting — that the activities through which identity is enacted are increasingly scaffolded, framed, or simulated by artificial systems, and that this scaffolding is not neutral with respect to the normative categories it underlies. If so, then the standard agenda of AI ethics, while indispensable, is incomplete. It addresses how artificial systems should be governed but leaves unexamined whether the subject on whose behalf governance is undertaken remains the kind of agent our theories assume. Whether this reconfiguration is best understood as loss, gain, or something more ambivalent is not settled here; the claim is only that it is significant enough that our normative categories must attend to it, in whichever direction it runs.

Each of the six functional domains identified here — intention formation, deliberation, memory, self-expression, preference formation, and relational interaction — warrants further empirical and philosophical investigation. So do the normative implications for responsibility, consent, dignity, and institutional design, particularly as they intersect with legal doctrine. And the relationship between the functional account offered here and existing debates in personal identity theory deserves fuller treatment. But the prior point is this: the question of identity under conditions of AI mediation is not speculative. It is a present philosophical problem. The background conception of the person that anchors our normative frameworks may have quietly changed, and that is a question philosophy cannot afford to defer.

Two clarifications about scope follow. The severity and distribution of these effects—how far, for whom, and under what conditions AI mediation reorganizes each function—are empirical questions that the present framework is intended to make tractable, and answering them will require sustained collaboration among philosophers, psychologists, and researchers who study the design and behavior of these systems. The contribution offered here is the prior, conceptual one: identifying which functions are at stake, why they are constitutive of identity, and what follows normatively if they are indeed being reorganized. The argument is therefore best read as a functional hypothesis that frames an interdisciplinary research program, not as an empirical estimate of its magnitude.

Data Access Statement

N/A

Contributor Statement

Michael A. Santoro — conceptualization, investigation, writing (original draft), writing (review and editing).

Use of AI

In preparing this manuscript, the author used Claude (Anthropic) for editorial support, confined to suggestions on phrasing and sentence-level expression. Every claim the manuscript makes, and every reason offered for it, rests on the author's own scholarly judgment. The assistant was not used to collect, analyze, or interpret evidence. Where it assisted in locating published sources, the author selected, read, and verified each one. The author has independently verified every cited source against the original publication, and takes sole responsibility for the content of the manuscript. No AI system is an author of this work, and no AI system meets the criteria for authorship.

Funding Statement

The author did not receive any financial support for the work on this article.

Acknowledgments

An early version of this argument was presented to a joint seminar of philosophers and data scientists at Santa Clara University, and a later version as an invited seminar jointly convened by the Institute for Assured Autonomy and the Department of Computer Science, Whiting School of Engineering, Johns Hopkins University. I am grateful to the participants in both seminars for questions that sharpened the argument.

Conflict of Interest Statement

There is no conflict of interest.

References

Braidotti, R. (2013). *The posthuman*. Polity.

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Coeckelbergh, M. (2022). Self-improvement: Technologies of the soul in the age of artificial intelligence. Columbia University Press.

Danaher, J. (2019). The ethics of algorithmic outsourcing in everyday life. In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (pp. 98–117). Oxford University Press.

Descartes, R. (1996). *Meditations on first philosophy* (J. Cottingham, Trans.). Cambridge University Press. (Original work published 1641).

Floridi, L. (2011). The informational nature of personal identity. *Minds and Machines, 21*(4), 549–566. https://doi.org/10.1007/s11023-011-9259-6

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

García-Vidal, F. J., Ciuti, C., & Ebbesen, T. W. (2021). Manipulating matter by strong coupling to vacuum fields. *Science, 373*(6551), eabd0336. https://doi.org/10.1126/science.abd0336

Hall, D. L., & Ames, R. T. (1987). *Thinking through Confucius*. State University of New York Press.

Haraway, D. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.

Heersmink, R. (2017). Distributed selves: Personal identity and extended memory systems. *Synthese, 194*(8), 3135–3151. https://doi.org/10.1007/s11229-016-1102-4

Heersmink, R. (2018). The narrative self, distributed memory, and evocative objects. *Philosophical Studies, 175*(8), 1829–1849. https://doi.org/10.1007/s11098-017-0935-0

Honneth, A. (1995). *The struggle for recognition: The moral grammar of social conflicts*. Polity.

Ihde, D. (1990). *Technology and the lifeworld*. Indiana University Press.

Kant, I. (1997). *Groundwork of the metaphysics of morals*. Cambridge University Press. (Original work published 1785).

Locke, J. (1996). *An essay concerning human understanding* (K. P. Winkler, Ed.). Hackett. (Original work published 1689).

Matei, S. (2024). Generative artificial intelligence and collective remembering: The technological mediation of mnemotechnic values. *Journal of Human-Technology Relations, 2*(1), 1–22. https://doi.org/10.59490/jhtr.2024.2.7405

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

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield International.

Pavanini, M. (2024). Postphenomenology and human constitutive technicity: How advances in AI challenge our self-understanding. *Journal of Human-Technology Relations, 2*(1), 1–25. https://doi.org/10.59490/jhtr.2024.2.7377

Santoro, M. A. (2026a, April 16). Designing lethal decisions: AI, accountability, and the future of military judgment. *Modern War Institute at West Point.* https://mwi.westpoint.edu/designing-lethal-decisions-ai-accountability-and-the-future-of-military-judgment/

Santoro, M. A. (2026b, March 30). Where is accountability when governments deploy AI? *Tech Policy Press.* https://www.techpolicy.press/where-is-accountability-when-governments-deploy-ai/

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.

Taylor, C. (1994). The politics of recognition. In A. Gutmann (Ed.), *Multiculturalism: Examining the politics of recognition* (pp. 25–73). Princeton University Press.

Tu, W. (1985). *Confucian thought: Selfhood as creative transformation*. State University of New York Press.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

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

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Wellner, G. (2024). A postphenomenological guide to AI regulation. *Journal of Human-Technology Relations, 2*(1). https://doi.org/10.59490/jhtr.2024.2.7397

Yaghoobian, H. (2025). The writer in-between: A post-phenomenological analysis of large language models (LLMs) and their implications for writer-tool relations. *Journal of Human-Technology Relations, 3*, 1–18. https://doi.org/10.59490/jhtr.2025.3.7398