

Digital Mediation, Selfhood, and Hybrid Agency: A Minimal Three-Dimensional Framework

Saida Agzamkhodjayeva

info@scientists.uz

National University of Uzbekistan named after Mirzo Ulugbek

Research Article

Keywords: digital mediation, extended cognition, digital selfhood, personal identity, hybrid agency, artificial intelligence

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10636563/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Claims that digital technologies blur the boundaries of consciousness and personal identity often collapse several distinct philosophical questions. This article develops a minimal three-dimensional framework of digital mediation that distinguishes cognitive integration, self-representational mediation, and agential distribution. Unlike existing multidimensional accounts that remain within one domain, the framework coordinates evidence across cognition, selfhood, and action while specifying which cross-domain inferences require an additional bridge argument. Digital artifacts may scaffold or partly constitute cognitive processes without extending phenomenal consciousness; profiles, avatars, and technologically organized memories may participate in narrative and practical selfhood without multiplying numerically distinct persons; and artificial systems may redistribute control within hybrid action without acquiring moral responsibility or personhood. Each dimension is assessed through explicit ordinal criteria rather than an impressionistic low–high scale. The framework is applied to conversational artificial intelligence, personal-memory systems, immersive avatars, recommender systems, and brain–computer interfaces. It also derives dimension-specific vulnerabilities and safeguards: continuity and provenance for cognitive dependence, contestability and contextual integrity for mediated selfhood, and meaningful control and responsibility mapping for distributed action. The result is a relational account of digital transformation that identifies how technologies reorganize thinking, self-interpretation, and control without positing a unitary “digital consciousness.”

1 Introduction

Digital technologies are now woven into remembering, planning, self-presentation, communication, and decision-making. A person may use a cloud calendar to organize future action, a conversational artificial intelligence system to formulate an argument, a social-media profile to maintain a public identity, an avatar to inhabit a virtual environment, and a recommender system to filter available options. These cases are frequently grouped under the claim that technology is changing, extending, or even digitizing human consciousness. The claim is rhetorically powerful but philosophically unstable. A calendar may be integrated into a cognitive routine without having any experience. An avatar may transform the way its user understands and presents herself without becoming a second person. An artificial system may make a difference to what a human–machine ensemble does without qualifying as a morally responsible agent. Treating these consequences as manifestations of one transformation obscures precisely what must be explained.

The difficulty is not simply terminological. “Consciousness,” “self,” “personal identity,” “personhood,” and “agency” identify different explananda. Phenomenal consciousness concerns what experience is like for a subject; cognition includes processes such as memory, inference, attention, and planning; numerical personal identity concerns whether someone at one time is the same individual as someone at another; narrative and practical selfhood concern how a life is interpreted, owned, and organized; agency concerns the production and control of action; and moral agency concerns the conditions under which an entity can properly be held responsible. These concepts overlap in human life, but no relation of

entailment runs automatically from one to another. Digital technology makes the need to separate them especially visible.

Three philosophical literatures illuminate different parts of this problem. Work on extended cognition asks whether external artifacts can be constituents of cognitive processes rather than merely causes or supports of them (Clark and Chalmers 1998; Clark 2008; Menary 2010). Work on narrative and distributed selfhood examines how autobiographical memory, social recognition, and material or digital objects participate in the constitution of a self (Schechtman 1996; Heersmink 2017, 2018; Bortolan 2024). Recent philosophy of artificial intelligence asks what kind of agency can be attributed to generative systems and how action is distributed in human–AI configurations (Floridi 2023; Barandiaran and Almendros 2025; Nyholm 2018). Each debate contains sophisticated internal distinctions, but discussions of “digital consciousness” commonly move between them without marking the transition. The result is conceptual inflation: a tool that changes cognition is described as altering consciousness; a mediated self-representation is treated as a new personal identity; or a causally effective artificial system is described as a person.

Existing approaches do not yet supply a common rule for coordinating these claims. Heersmink’s multidimensional analysis explains degrees of integration within embedded and extended cognitive systems (Heersmink 2015), but its dimensions do not determine whether a technologically mediated relation is identity-forming or agential. Postphenomenology explains how technologies mediate relations between subjects and worlds (Verbeek 2005, 2008), but does not itself provide a comparative scheme for separating cognitive dependence, self-constitution, and control. Accounts of artificial and collaborative agency clarify how action can be distributed (Floridi 2023; Nyholm 2018), yet they do not show when the same system also becomes part of autobiographical self-interpretation. The unresolved question is therefore: **how can technologically mediated changes in cognition, selfhood, and action be compared without treating evidence in one domain as sufficient evidence in another?**

This article answers that question with a cross-domain framework. It uses “digital mediation” as a neutral term for technologically organized relations that alter how a person engages with the world, with herself, or with courses of action. The framework contains three minimal analytically independent dimensions:

1. **Cognitive integration:** the degree to which a digital artifact is reciprocally and reliably incorporated into a cognitive process.
2. **Self-representational mediation:** the degree to which a digital artifact participates in the interpretation, presentation, and practical organization of the person’s self.
3. **Agential distribution:** the degree to which the selection, control, and execution of action are spread across human and artificial components.

They are minimal because each answers a different question, is supported by a different evidential set, and predicts a different primary vulnerability. Cognitive integration concerns the constitution of information-processing routines and predicts dependence on continuity and reliability. Self-

representational mediation concerns autobiographical interpretation and social recognition and predicts identity-related vulnerability. Agential distribution concerns the allocation of practical control and predicts responsibility and oversight problems. Phenomenal consciousness is treated as a boundary question rather than a fourth score because none of these relational indicators is, by itself, evidence of phenomenal realization. Cultural and institutional conditions modify all three dimensions instead of constituting a single homogeneous fourth dimension.

The dimensions are graded rather than binary and analytically independent rather than mutually exclusive. A technology can score highly on one and weakly on another. A brain–computer interface may be tightly integrated into a cognitive operation while remaining peripheral to its user’s public identity. An avatar may be central to self-interpretation while playing only a modest role in reasoning. A partly autonomous artificial-intelligence agent may significantly influence action while having no credible claim to phenomenal consciousness. Section 6 replaces intuitive scoring with explicit low, medium, and high criteria and requires every assessment to be justified at the level of a particular user, task, and period of use.

The framework supports three negative theses and one positive thesis. The negative theses are that cognitive extension does not by itself entail extended phenomenal consciousness; self-representational multiplication does not by itself entail the multiplication of persons; and distributed agency does not by itself entail artificial personhood or moral responsibility. The positive thesis is that digital technologies transform persons by reconfiguring relations of cognitive dependence, self-interpretation, and control. This relational transformation can be profound even when the dramatic ontological conclusions are rejected.

The contribution is therefore not a new criterion of consciousness, a complete metaphysics of personal identity, or a universal theory of agency. It is a minimal coordinate framework with four commitments: domain-specific criteria, explicit constraints on inference, bridge conditions for stronger conclusions, and a proportional mapping from degrees of mediation to types of vulnerability. This distinguishes the proposal from a merely descriptive taxonomy. It explains why loss of access to a deeply integrated memory system may be a cognitive harm, manipulation of an avatar or profile may be an identity-related harm, and opaque automation may create a control and accountability problem. These harms arise through different mechanisms and therefore require different remedies.

The argument proceeds as follows. Section 2 establishes the distinctions between cognition and phenomenal consciousness, numerical identity and selfhood, and functional agency and moral agency. Sections 3–5 analyze cognitive integration, digital selfhood, and hybrid agency in turn. Section 6 formalizes the three-dimensional framework and applies it to five technological configurations. Section 7 answers five objections. Section 8 develops epistemic, ethical, and design implications. The conclusion explains what digital mediation changes—and what it does not.

2 Three Distinctions That Digital Philosophy Must Preserve

2.1 Cognition is not synonymous with phenomenal consciousness

The first distinction concerns the scope of the mental. Philosophical and scientific discussions use “consciousness” for several phenomena. Block’s distinction between access consciousness and phenomenal consciousness remains useful here: information may be available for reasoning, report, and the guidance of action without settling whether there is something it is like to undergo the relevant state (Block 1995). The same caution applies to cognition more generally. Memory retrieval, inference, categorization, and planning can be investigated in functional terms, while phenomenal consciousness concerns subjective experience and the explanatory problem associated with it (Chalmers 1996). A theory that locates some cognitive processing beyond the brain therefore does not automatically locate experience beyond the brain.

The extended-mind thesis is primarily a claim about the constitution of cognitive processes. In the original parity-based argument, a reliably available external resource can play a role equivalent to an ordinary internal memory process (Clark and Chalmers 1998). Later work develops a broader picture in which cognition is realized by coordinated neural, bodily, and environmental processes (Clark 2008; Menary 2010). This thesis challenges intracranialism about cognition. It does not, without a further argument, establish that a notebook, smartphone, server, or language model is part of the physical machinery of phenomenal experience. Clark explicitly resists this inference, arguing that the considerations supporting extended cognition do not straightforwardly generalize to the machinery of consciousness (Clark 2009).

This point matters because digital systems often possess features that make them cognitively powerful while remaining evidentially irrelevant to phenomenal consciousness. A language model can make information available, produce candidate inferences, and support problem solving. Those functions may alter the organization of a user’s cognitive activity. Whether the model itself is conscious is a different question, requiring a theory of consciousness and evidence that the architecture satisfies its relevant conditions. Contemporary assessments of artificial consciousness accordingly examine recurrent processing, global availability, higher-order representation, predictive processing, agency, and other proposed indicators rather than inferring consciousness from fluent output or functional usefulness (Chalmers 2023; Butlin et al. 2023). The first rule of the present framework follows: evidence of cognitive contribution should be interpreted as evidence about cognition unless an independent bridge to phenomenal consciousness is supplied.

2.2 Numerical personal identity is not the same as narrative or practical selfhood

The second distinction separates several questions often compressed into the phrase “personal identity.” Numerical identity asks whether an individual existing at one time is one and the same individual as an individual existing at another. Classical debates address bodily, psychological, and other

relations that might ground this persistence. Parfit's reductionist treatment emphasizes psychological continuity and connectedness while also questioning whether identity itself has the practical importance commonly assigned to it (Parfit 1984). Whatever account is accepted, numerical identity is a one-to-one relation: if an online profile and its user are not literally the same entity, similarity or representation alone does not make them numerically identical.

Questions of characterization and practical selfhood are different. They concern who someone takes herself to be, which experiences she appropriates as her own, which commitments organize her life, and how she is recognized by others. Narrative approaches maintain that self-interpretation is not an optional description placed over an otherwise complete life. The ability to connect experiences into an intelligible and practically coherent narrative is partly constitutive of the kind of self whose actions, projects, and relationships unfold over time (Schechtman 1996, 2014). These claims do not require a narrative to be perfectly unified, transparent, or authored in isolation. Narratives are socially scaffolded and materially supported.

Digital environments intervene directly in this second domain. Profiles, archives, images, avatars, and recommendation histories store traces, solicit self-descriptions, present past actions to audiences, and shape possible future performances. They can therefore alter self-understanding and social recognition. Yet a transformation in selfhood is not automatically a change in numerical identity. A person may sustain multiple context-sensitive profiles without becoming numerically many persons. Conversely, a profile can be central to a person's practical identity even though it is not itself a person. This is not a deflationary conclusion. It locates the genuine importance of digital self-representation in the organization of a life rather than in an unsupported multiplication of substances or subjects.

Personhood must also be distinguished from both numerical identity and selfhood. "Person" can name a metaphysical category, a moral status, or a legal status. Capacities such as consciousness, rationality, autonomy, responsibility, and participation in normative relations enter different theories of personhood in different combinations. A digital representation may transform personal identity in the narrative sense without acquiring any such status. Thus, evidence that an avatar matters to who someone is does not establish that the avatar is a separate bearer of rights.

2.3 Functional agency is not yet moral agency

The third distinction concerns agency. At its thinnest, agency involves goal-directed sensitivity and a capacity to make a difference to the production of action. Richer accounts require autonomous goal formation, reasons-responsiveness, understanding, self-maintenance, or reflective control. Moral agency adds normative competence and an appropriate basis for responsibility. Because these notions form a family rather than a single threshold concept, claims about artificial agency must specify the sense in which "agent" is being used.

Generative artificial intelligence makes this specification urgent. A conversational model can propose plans, revise text in response to feedback, select tool calls, and produce outputs that redirect a user's

conduct. In a suitably organized system, artificial components can contribute to goal pursuit in ways that are neither passive nor accurately described as the mere transmission of a fixed human intention. Floridi therefore argues for artificial agency without attributing intelligence in the full human sense (Floridi 2023). Barandiaran and Almendros (2025), by contrast, examine the architecture and mode of operation of large language models to distinguish their agent-like capacities from autonomous living agency. These accounts differ, but both show why an all-or-nothing contrast between “mere tool” and “person” is inadequate.

The concept of hybrid agency captures an intermediate and relational phenomenon. In human–AI activity, goals may originate with a person, options may be generated by a model, selection may be negotiated through prompts and revisions, execution may be delegated to software, and monitoring may be shared across interfaces and institutions. The resulting action depends on a system of contributions. Describing this distribution does not settle whether any artificial component has intentions, phenomenal consciousness, moral understanding, or responsibility. Floridi and Sanders’s influential account of artificial moral agents shows one way of widening the concept of moral agency (Floridi and Sanders 2004), while Véliz (2021) argues that algorithmic systems lack the capacities required for moral agency and should be treated as “moral zombies.” The disagreement cannot be resolved by noting that an AI output had causal effects. Causal contribution, functional agency, and moral accountability must be assessed separately.

The three distinctions yield three prohibited inferences:

- from integration into cognition to participation in phenomenal consciousness;
- from the multiplication of self-representations to the multiplication of persons;
- from contribution to action to moral agency or personhood.

The remainder of the article shows that rejecting these inferences does not minimize the transformative power of digital technologies. On the contrary, it allows that power to be described with greater precision.

3 Cognitive Integration Without Automatic Consciousness Extension

3.1 From external support to integrated process

The extended-mind debate begins from a challenge to a location-based prejudice: why assume that every constituent of cognition must be inside the skull? Clark and Chalmers’s well-known case of Otto and his notebook asks readers to compare biological memory with a reliably consulted external record (Clark and Chalmers 1998). The point is not that every object carrying information is part of a mind. It is that an external resource may, in the right recurrent organization, perform a role within the process by which a person remembers and acts.

Critics respond with the coupling–constitution objection. The fact that an internal cognitive process is causally coupled to an external artifact does not show that the artifact constitutes the process; digestion depends on external conditions without incorporating every cause into the digestive system. Adams and Aizawa (2001) additionally argue that cognitive processes possess distinctive properties, including forms of non-derived content, that ordinary external vehicles lack. The objection blocks an indiscriminate extension of cognition to everything that influences thought.

A productive response is to replace a single parity test with a multidimensional account of integration. Heersmink (2015) identifies factors such as information flow, reliability, durability, trust, procedural transparency, personalization, and transformation. These factors explain why a frequently used and personalized artifact can become deeply integrated while an occasionally consulted website remains an external aid. The dimensions are not individually necessary and jointly sufficient conditions for constitution. They offer an empirically sensitive profile of the relation between agent and artifact.

This graded approach is especially appropriate for digital technology. Cloud calendars are persistently available but depend on infrastructure outside the user’s control. Search engines supply vast information but are weakly personalized and require repeated interpretation; internet-based cognition is therefore best analyzed as a heterogeneous ecology rather than as a single extension (Smart 2017). A brain–computer interface may establish rapid and reciprocal information flow while remaining opaque or unstable. Heersmink’s analysis of brain–computer interfaces accordingly distinguishes embodied tools and cognitive tools rather than treating physical proximity as decisive (Heersmink 2013). What matters is the organization of activity: how information moves, how the person has learned to rely on the artifact, how easily the resource can be accessed, and whether the combined system displays a stable capacity.

The first dimension of digital mediation—cognitive integration—should therefore be understood as the degree to which a technological resource participates in a recurrent cognitive routine. High integration involves more than causal influence. The interaction is relatively reliable and durable; the resource is available when needed; information flows in both directions; the user is trained or habituated in its use; the resource is personalized to the user or task; and the combined process produces capacities that neither component exercises in the same way alone. Failure on one factor need not eliminate integration, but it should reduce confidence in a constitutive claim.

3.2 Large language models as a hard case

Large language models are a particularly demanding test because they are neither stable notebooks nor ordinary search engines. They respond dynamically to natural-language input, generate novel sequences, and can be incorporated into iterative problem solving. A user may ask a model to reorganize a plan, challenge a premise, supply counterexamples, or preserve a developing record of a project. Smart, Clowes, and Clark (2025) argue that such interactions can support forms of extended cognition. Kudina, Ballsun-Stanton, and Alfano (2025) describe language models as scaffolds for proleptic reasoning: users can engage with a model in order to reach forms of reasoning they are still learning to perform independently.

The strongest version of the extension claim nevertheless applies only to some user–system relations. A first-time query followed by uncritical copying shows access to an external service, not a deeply integrated cognitive process. By contrast, a long-term workflow in which the user supplies domain knowledge, evaluates responses, develops personalized prompt practices, maintains a project history, and repeatedly revises outputs may exhibit reciprocal and transformative integration. The relevant unit is not “the language model” in abstraction but a temporally extended system consisting of a person, interfaces, stored context, institutional rules, and model services.

Two features of language models limit integration. First, their operations are epistemically opaque to ordinary users. A fluent answer does not reveal which internal or external sources produced it, and the phenomenology of smooth dialogue can invite excessive trust. Heersmink et al. (2024) therefore emphasize transparency, trust, and trustworthiness as separate issues. A system can feel transparent in use while remaining opaque in the evidential sense. Second, outputs are variable and can be inaccurate. Reliability is not a fixed property of a model; it depends on the domain, the interface, retrieval mechanisms, the user’s verification practices, and the costs of error. A responsible cognitive partnership must include checking and correction rather than treating linguistic confidence as epistemic authority.

These limitations support a conditional conclusion. Language models can be cognitively integrated, but their integration is task-relative, user-relative, and revisable. “Scaffold” is often the better description when the system enables performance without becoming a stable constituent of the user’s cognitive routine. “Extension” becomes plausible when reciprocal reliance, personalization, availability, and transformation are sufficiently strong. The framework does not settle the coupling–constitution debate by definition; it locates the facts on which that debate should turn.

3.3 Why cognitive integration does not settle consciousness

Even a high score on cognitive integration provides no shortcut to phenomenal consciousness. The mistake can take two forms. One is to infer that because an external resource is part of a person’s cognitive system, it is part of the physical basis of that person’s experience. The other is to infer that because an artificial system performs cognitive functions, it has experiences of its own. Both conclusions require additional premises.

The first inference is resisted by Clark’s position that extended cognitive machinery need not imply extended conscious machinery (Clark 2009). Cognitive processes can unfold over longer time scales and through heterogeneous media. Phenomenal consciousness may depend on forms of high-bandwidth reciprocal processing or biological organization not present in ordinary artifact use. The present framework remains neutral on whether consciousness could ever extend beyond the brain. Its narrower claim is that the evidence establishing cognitive integration—reliability, access, personalization, reciprocal information flow—does not establish the presence or spatial extent of phenomenal experience.

The second inference confuses performance with subjectivity. Contemporary language models can generate reports about emotions or consciousness, but a first-person sentence is not independent evidence of a first-person point of view. Chalmers (2023) argues that current systems face significant obstacles under several leading theories of consciousness while leaving open the possibility that future systems may overcome them. Butlin et al. (2023) likewise develop theory-derived indicator properties for assessing artificial consciousness. Such work illustrates the proper evidential route: begin with candidate theories of consciousness, derive architecture-relevant conditions, and evaluate the system. Cognitive usefulness to a human is not one of those conditions by itself.

The conclusion is therefore asymmetric. Digital artifacts can substantially reconfigure human cognition even if neither the artifact nor the extended system supports a new locus of phenomenal experience. The transformation is real, but it belongs initially to the organization of cognition.

4 Digital Selfhood Without Multiplying Persons

4.1 The self as technologically mediated

Digital artifacts do not only support problem solving. They also participate in how people answer practical questions about who they are. Early accounts of online identity already emphasized that networked environments allow users to experiment with multiple presentations of self (Turkle 1995), while the contemporary infosphere makes such mediation pervasive rather than exceptional (Floridi 2014). Social-media profiles, message histories, photographs, quantified-self records, game avatars, and personal archives present selected aspects of a life to the person and to others. These artifacts are not neutral containers. Their interfaces solicit categories, rank memories, stabilize some descriptions, hide others, and make past performances available for future interpretation. They are technologies of self-relation.

Narrative theories of selfhood provide a framework for understanding this role. On Schechtman's view, a person constitutes herself as a self by organizing experiences into an intelligible narrative constrained by reality and by the requirements of interpersonal articulation (Schechtman 1996). Her later person-life view stresses the practical unity of a life without reducing personal identity to a single psychological relation (Schechtman 2014). Digital mediation matters because it changes the materials, audiences, and temporal structure of self-narration.

The idea that selfhood can be distributed across artifacts is already present in work on memory and identity. Heersmink (2017) argues that extended memory systems can contribute to personal identity, and his analysis of evocative objects shows how material artifacts scaffold autobiographical narratives (Heersmink 2018). His later account of narrative niche construction explains how memory ecologies can stabilize and transform distributed narrative identities over time (Heersmink 2020). Digital environments intensify these relations. They can preserve large searchable archives, automatically resurface events, connect records to public responses, and permit continual revision of a profile. Bortolan (2024) argues

that online activity can contribute to narrative self-constitution rather than merely display a pre-existing self. The online world is therefore neither a detached theatre nor a perfect mirror. It is one of the environments in which selves are enacted and interpreted.

This account should not universalize an individualist picture of self-authorship. Confucian role ethics treats a person's practical identity as constituted through lived roles and cultivated relations rather than as the expression of an antecedently self-sufficient individual (Ames 2011). Menkiti's influential, though contested, communitarian account of African personhood likewise emphasizes social incorporation and recognition in the formation of personhood (Menkiti 1984). These traditions differ from narrative theories and from one another, but they expose a common limitation in purely profile-centered analysis: digital mediation reorganizes not only what an individual says about herself, but also the relational practices through which roles, standing, and recognition are conferred. The self-representational dimension is therefore culturally and institutionally variable. Its degree cannot be inferred from the quantity of personal data or intensity of self-expression alone; it depends on the significance of the mediated relation within a person's social world.

Four relations are especially important. First, a digital representation can be **expressive**: the person uses it to articulate commitments, affiliations, and aspirations. Second, it can be **mnemonic**: it organizes access to autobiographical materials and thereby affects what is available for self-interpretation. Third, it can be **recognitional**: others respond to the profile or avatar, and those responses feed back into the person's social identity. Fourth, it can be **practical**: the representation opens and closes options, sustains projects, and guides future conduct. When these relations become durable and affectively salient, the artifact is not merely a sign pointing to a self located elsewhere. It participates in the ongoing constitution of practical and narrative selfhood.

4.2 Profiles, avatars, and digital selves

The language of a "digital self" can now be interpreted in at least three ways. A deflationary view treats profiles and avatars as representations: they are about a person but not parts of the person. A constitutive view holds that some digital structures participate in the relations through which the self is maintained. A stronger realist view treats certain avatars, profiles, or AI replicas as genuine extensions or components of the self. Declos and Grandjean (2026) defend a version of this realist position and explicitly connect digital selves to debates about psychological continuity, virtual reality, and the extended mind.

The three-dimensional framework can accommodate the insight behind stronger views without treating all digital representations alike. A dormant account containing generic information is weakly connected to selfhood. A long-used avatar through which a person sustains relationships, develops skills, and undertakes recognized commitments can be strongly self-mediating. An AI-generated replica trained on personal data may resemble the user and interact with others, yet resemblance is insufficient for incorporation. The relevant questions concern autobiographical salience, affective investment, public

recognition, practical consequence, and the person's capacity to endorse or contest what the artifact presents.

This last condition is important because self-mediation can be alienating as well as enabling. Platform templates may force a person into categories she rejects. Recommendation systems may infer interests that become visible to others and eventually to the person herself. Automated memories may privilege what was frequently photographed rather than what was meaningful. A profile can continue to circulate an obsolete self-description. High self-representational mediation therefore does not imply authentic self-expression. It identifies a degree of influence over the organization of selfhood, which may be supportive, distorting, or coercive.

Virtual-reality avatars make the distinction between selfhood and numerical identity especially clear. Embodied interaction through an avatar can alter perception, social confidence, and the range of roles a person can explore. The avatar may become a stable site of identification. Nevertheless, neither embodiment nor identification alone creates a numerically distinct person. The same individual can inhabit multiple avatars, just as one individual can occupy several social roles. If an avatar were to acquire independent consciousness, memory, and autonomous practical projects, the metaphysical question would change. But that conclusion would follow from new facts about the entity, not from visual embodiment or representational multiplicity.

4.3 AI-curated memory and narrative authority

The integration of language models with personal archives creates a more complex case. A conventional archive stores materials that a person or platform later retrieves. An AI-enabled memory system can solicit information, summarize episodes, connect records, answer autobiographical questions, and generate narratives. Smart, Clowes, Krueger, and Boniface (2026) argue that such systems can become active participants in personal memory rather than passive repositories. Their role links cognitive integration to self-representational mediation: the same system can support recall and influence the story through which recalled events become intelligible.

This overlap makes the independence of the dimensions more, not less, important. A system may be an effective retrieval aid while exerting little influence on identity. It may also be only moderately reliable as a memory aid while strongly shaping self-interpretation through persuasive summaries. The risks differ. False recall is an epistemic failure; the imposition of a reductive life story is an identity-related failure; unauthorized disclosure is a privacy failure. A single label such as "extended self" does not distinguish them.

Narrative authority should therefore remain contestable. A digital system may propose connections, but the person must be able to inspect sources, reject summaries, preserve alternative interpretations, and understand when a narrative is inferred rather than recorded. This requirement does not assume a sovereign self who authors a life without social influence. Human selfhood is always relational and

scaffolded. It requires only that technological mediation not be mistaken for an infallible discovery of who the person “really” is.

The conclusion of this section is twofold. Digital artifacts can be constitutive of narrative and practical selfhood when they participate in expression, memory, recognition, and action. Yet this constitutive role does not multiply numerical persons. The transformation concerns the mediated unity, intelligibility, and social presentation of a life.

5 Hybrid Agency Without Artificial Personhood

5.1 Action in a human–AI loop

Digital systems increasingly do more than present information. They rank possibilities, generate plans, initiate transactions, and adapt outputs to feedback. It is tempting to describe them either as neutral tools controlled by a human or as autonomous agents replacing human control. Both descriptions can be misleading. In many cases, action emerges from a loop in which different agential functions are allocated to different components.

Consider a researcher working with a conversational AI system. The human initially frames a question and supplies constraints. The model generates possible structures and objections. The human selects, rejects, or modifies them. The system drafts language; the human checks sources and revises claims. Institutional policies determine which uses are permitted, and interface design determines what information is visible. If the system can call external tools, it may also execute searches, schedule tasks, or transform files. The final action cannot be explained by pointing only to the initial human intention or only to the artificial output. It is produced by an organized sequence of contributions.

The same structure appears outside writing. A navigation system proposes a route in response to a destination set by the driver, continuously recalculates in light of traffic data, and changes the driver’s movement. A recommender system orders options and thereby alters what the user is likely to choose. A clinical decision-support system identifies patterns, but a professional interprets the recommendation within a normative practice. An AI agent may divide a general task into subgoals and call tools, while a user retains the power to authorize or interrupt execution. The degree of agential distribution varies according to which functions are delegated and how effectively the human can monitor and reverse them.

Five functions are useful for analysis:

- **goal formation:** determining what outcome is to be pursued;
- **option generation:** constructing candidate ways of reaching the goal;
- **selection:** choosing among available options;
- **execution:** producing changes in the world or in an informational environment;

- **monitoring and correction:** evaluating results and revising the course of action.

No single allocation of these functions defines hybrid agency. A system can generate options without selecting them, or execute a selected plan without understanding its purpose. The agential profile becomes more distributed as artificial components occupy more functions, exercise wider discretion within them, and become less continuously supervised.

5.2 Agency in a thin but non-trivial sense

The language of agency is justified when it marks organized sensitivity to goals and conditions rather than merely causal impact. A stone contributes causally to a broken window but does not select or adapt its behavior. To count as an agential participant in the present framework, an artificial component must satisfy four diagnostic conditions within a particular action loop: it must (1) operate under an objective or action-guiding constraint; (2) select or construct alternatives rather than transmit a fully specified command; (3) alter its output in counterfactual response to relevant changes in the task or environment; and (4) have an authorized output that controls, or materially structures, a subsequent stage of action. These conditions distinguish agential participation from mere causal influence without presupposing consciousness, intrinsic goals, or moral understanding.

The conditions are functional and relational, not a proposed essence of agency. A spam filter may satisfy them within a narrow domain, while a language model used only to correct punctuation may not. Higher agential distribution occurs when the artificial component occupies several functions, has a wider discretionary range, sets subgoals or chooses means, and proceeds with less moment-to-moment supervision. This account is consistent with Nyholm's distinction between domain-specific, supervised machine agency and responsible human agency in human-machine collaborations (Nyholm 2018). Floridi's account of artificial agency without intelligence similarly captures the possibility that successful adaptive action need not reproduce human understanding (Floridi 2023).

Thin agency should not be confused with a claim that the system is an autonomous living agent. Barandiaran and Almendros (2025) distinguish the operation of large language models from the self-maintaining organization emphasized by enactive accounts. A model's apparent initiative depends on training processes, prompts, computational infrastructure, and the architectures wrapped around it. Its goals are typically assigned or engineered, and its activity does not exhibit the same intrinsic normativity as an organism's continued self-production. Nevertheless, once integrated into an agentic architecture, the system can transform the mode of human action. It changes which possibilities are generated, how quickly they can be evaluated, and how far execution can proceed without moment-to-moment human intervention.

This is why "tool" and "agent" should not be treated as exclusive natural kinds. A system can be a designed instrument within a broader practice and an agent-like contributor within a particular action loop. Verbeek's analysis of technological mediation supports this relational perspective. Technologies do not simply stand between a pre-formed subject and a stable world; they help shape intentional

relations and possible actions (Verbeek 2005, 2008). Aydin, González Woge, and Verbeek (2019) extend this insight by describing technology as a mediating milieu. Hybrid agency names one consequence of such mediation: practical intentionality is reorganized across a technologically structured relation.

5.3 Responsibility is not dissolved by distribution

Once agency is described as distributed, a predictable worry arises: if no single component produces the entire action, perhaps no one is responsible. That conclusion does not follow. Causal distribution can complicate responsibility, but it does not remove institutional and human obligations. Responsibility depends not only on which component generated an output but also on who selected the system, defined the task, established oversight, possessed relevant competence, benefited from the action, and had the power to prevent or correct harm.

The distinction between functional and moral agency is decisive. Floridi and Sanders (2004) propose a relatively inclusive account on which artificial agents may qualify as sources of morally significant action at an appropriate level of abstraction. Véliz (2021) argues that algorithms lack consciousness, understanding, intentions, and moral reasons, and therefore cannot be moral agents in the responsibility-bearing sense. The present framework need not resolve the entire dispute. It requires only that the moral conclusion be separately defended. A system's participation in option generation or execution shows agential distribution; it does not show the capacities that a chosen theory regards as necessary for blame, obligation, or personhood.

In ordinary present-day cases, the safer attribution is that human actors and institutions remain accountable for governing a hybrid system. Their responsibility may be differentiated: developers answer for foreseeable design choices, deploying institutions for selection and oversight, and professional users for competent use within their roles. Nyholm's analysis directs attention to the human loci that initiate, supervise, and manage collaborative agency (Nyholm 2018). Santoni de Sio and van den Hoven (2018) add a more demanding account of meaningful human control: system behavior should track relevant human moral reasons, and outcomes should be traceable to agents who understand their role and can be called to account. These requirements are stronger than placing a nominal approver at the end of an opaque process.

The artificial component can still be causally and functionally significant. Indeed, recognizing its contribution helps identify where control has migrated and where new oversight is needed. Treating the system as a mere passive tool can hide this migration; treating it as a responsible person can allow accountable humans to hide behind it.

The appropriate conclusion is therefore neither human exclusivism about causal agency nor artificial personhood. Digital systems can redistribute the production and control of action. This hybridization is philosophically important because it changes the conditions under which human agency is exercised. Moral and legal responsibility must be assigned through an additional normative analysis.

6 A Three-Dimensional Framework of Digital Mediation

6.1 Minimality, criteria, and bridge conditions

The preceding analyses can now be integrated. Let a technologically mediated configuration at a particular time be represented as an ordinal profile:

$$M = \langle C, S, A \rangle$$

where **C** is cognitive integration, **S** is self-representational mediation, and **A** is agential distribution. The bearer of the profile is not a device type but a specified person–technology relation during a specified practice and period. The same application may therefore occupy different positions for different users, tasks, and stages of adoption.

The three dimensions are minimal in a methodological sense. Each corresponds to a distinct explanandum: the constitution of a cognitive process, the mediation of self-interpretation and recognition, or the distribution of practical control. Each also has a distinct evidential base and characteristic failure mode. Adding more device features would improve description but would not create a fourth explanandum at the same level. Phenomenal consciousness remains a separate boundary question because determining its realization requires a theory of consciousness. Cultural norms, platform governance, and institutional power are cross-cutting conditions: they change the degree and significance of C, S, and A rather than collapsing into one comparable score.

The terms low, medium, and high are reasoned ordinal categories, not measurements. Table 1 states the diagnostic criteria used to assign these categories across the three dimensions.

Table 1
Diagnostic criteria for the three dimensions of digital mediation

Dimension	Low	Medium	High
Cognitive integration (C)	Episodic, one-way, and readily replaceable assistance; interruption causes little loss of established capacity	Recurrent and reasonably reliable support with some learned reliance or personalization; interruption slows or complicates performance	Reciprocal, durable, readily available, personalized, and disruption-sensitive incorporation into a routine; loss impairs an established cognitive capacity
Self-representational mediation (S)	Generic or peripheral representation with little autobiographical, recognitional, or practical significance	Context-specific representation that influences self-description, audience recognition, or some continuing projects	Durable autobiographical and social authority across important roles or commitments; manipulation, exclusion, or loss threatens practical identity or recognition
Agential distribution (A)	Execution of a fixed human specification with no meaningful construction or selection of alternatives	Generation or narrowing of options under active human selection and authorization	Discretionary selection of means or subgoals plus execution or monitoring with limited real-time human supervision and significant effects on the course of action

An assessment should use the lowest category whose defining conditions are fully supported and should state the evidence and temporal scope. “High” cannot be inferred from frequency of use, emotional attachment, technical complexity, or unpredictability alone. For C, disruption sensitivity must concern an acquired capacity rather than inconvenience. For S, public visibility is insufficient without autobiographical, recognitional, or practical significance. For A, causal influence is insufficient without objective-guided selection and counterfactual adaptation.

The framework also specifies bridge conditions for stronger conclusions. Moving from C to a claim about phenomenal consciousness requires a theory-derived account of phenomenal realization and architecture-relevant evidence. Moving from S to numerical identity or personhood requires an account of persistence or status that explains why a representation is the same person or an independent bearer of standing. Moving from A to moral agency requires normative competence, reasons-responsiveness, and an account of answerability. These bridge requirements are defeasible research obligations, not declarations that the stronger conclusions are impossible.

The dimensions can interact through causal and developmental mechanisms. Repeated cognitive reliance can make a system autobiographically salient; a socially authoritative profile can shape goal formation; and an AI-curated memory can combine recall support, narrative framing, and option generation. Such interactions explain why digital transformations often feel unified in experience.

Analytic independence does not deny interaction; it prevents evidence of interaction from being treated as identity among the dimensions.

6.2 Comparative profiles

The framework can be applied to representative configurations. Table 2 presents conditional profiles justified by the criteria above, not judgments about every product in a category.

Table 2
Comparative profiles of representative technological configurations

Configuration	Cognitive integration	Self-representational mediation	Agential distribution	Principal philosophical issue
Conversational AI used intermittently for drafting	Low–medium: useful scaffolding but limited reliability and personalization	Low: little autobiographical or public significance	Medium: generates options, while the user selects and authorizes	Avoid equating fluent assistance with understanding or authorship
Personalized LLM-based memory system	High when reliable, persistent, and embedded in recall routines	High when it curates autobiographical narratives	Medium: prompts, summarizes, and frames retrieval without necessarily setting life goals	Preserve provenance, contestability, and narrative authority
Immersive social avatar	Medium: embodied interaction may reorganize attention and skill	High when central to relationships, commitments, and recognition	Medium: action is mediated through avatar and platform rules	Distinguish identity transformation from multiplication of persons
Brain–computer cognitive interface	High when reciprocal, transparent in use, and stable	Low–medium, depending on incorporation into self-understanding	Variable, rising with adaptive or automatic control	Distinguish bodily proximity from cognitive constitution and responsibility
Algorithmic recommender or decision system	Low–medium as a cognitive aid	Medium when repeated rankings shape preferences and self-description	Medium–high when it structures options and defaults	Identify hidden transfers of control without attributing moral personhood

Several comparisons follow. First, the avatar and the brain–computer interface reverse a common expectation. The technologically intimate interface may be highly cognitive but only modestly identity-mediating; the visually external avatar may be central to selfhood. Second, the personal-memory system can be high on both cognition and selfhood while remaining only moderately agential if the user retains control over goals and endorsement. Third, an algorithmic recommender can have a high agential impact

despite low cognitive integration: it shapes the field of action from the environment rather than becoming part of a recurrent process experienced as the user's own.

These cases also show why device-level classification is inadequate. The question is not whether "AI," "virtual reality," or "brain-computer interfaces" extend the mind. It is whether a specific person-technology relation, organized in a specific practice, satisfies relevant conditions. Integration and mediation are achieved, maintained, and sometimes lost. A software update can reduce reliability; a change in platform governance can undermine control; a once-central avatar can become autobiographically irrelevant; a new memory function can convert a generic assistant into an identity-shaping archive.

6.3 Explanatory and normative payoffs

The framework generates more than labels. It predicts different forms of vulnerability. High cognitive integration creates dependence: loss, manipulation, or interruption of the resource can impair an established cognitive capacity. High self-representational mediation creates exposure to identity-related harms: imposed categorization, decontextualized resurfacing, impersonation, or deletion can disrupt how a person is recognized and how she understands her life. High agential distribution creates control and accountability risks: the person may no longer know which component selected an option or may be unable to interrupt execution.

The framework also identifies mismatches. A user may experience a system as cognitively transparent and "part of me" while the system is institutionally controlled by a provider. A profile may be highly identity-mediating while the person has little power to correct it. An organization may delegate option generation and execution to AI while presenting a nominal human approval as sufficient oversight. Such mismatches are philosophically significant because subjective incorporation, functional integration, and governance can diverge.

The normative step is not that every high score automatically creates a right. It proceeds through foreseeable dependence and asymmetric control. When designers or institutions invite a person to organize an important cognitive capacity, social identity, or course of action around a system, retain control over the conditions of access and operation, and can reasonably mitigate resulting risks, they acquire stronger *pro tanto* duties of protection. The degree and type of mediation identify the protected interest and the foreseeable mode of harm; ordinary considerations of control, benefit, and capacity to prevent harm identify who bears the duty.

This argument supports proportional and dimension-specific safeguards. High C strengthens the case for continuity, reliability, exportability, and provenance because interruption or corruption can impair an acquired capacity. High S strengthens the case for access, correction, contextual integrity, and protection against impersonation because the system participates in recognition and practical identity. High A strengthens the case for auditability, reversibility, meaningful human control, and advance allocation of responsibility because discretionary functions have migrated into a hybrid process. The

protections are cumulative when a system is high on more than one dimension, as in an AI-curated personal-memory system. None depends on declaring the technology conscious.

7 Objections and Replies

7.1 Objection: the framework is only a taxonomy

One might object that the three dimensions merely rename familiar debates. A taxonomy can organize cases, but it does not explain why digital technologies transform persons or resolve any metaphysical question.

The objection would be decisive if the framework only placed examples into boxes. Its contribution becomes clearer by comparison. Heersmink (2015) offers multiple dimensions *within C*: reliability, durability, trust, personalization, and related features describe the integration of cognitive systems. The present framework retains those insights but does not relabel them as three new dimensions. It asks whether the same technology–person relation is also identity-mediating or agential, questions that Heersmink’s integration profile cannot answer. Postphenomenology supplies the general thesis that technology mediates subject–world relations (Verbeek 2005, 2008); the present framework adds a rule for disaggregating three kinds of mediation and comparing their evidential and normative consequences. Accounts of artificial agency explain domain-specific or collaborative machine action (Floridi 2023; Nyholm 2018); the present framework locates that action alongside, but not inside, cognitive extension and selfhood.

The resulting proposal makes four substantive commitments. First, it gives dimension-specific inclusion criteria rather than device labels. Second, it imposes bridge conditions: a conclusion in one domain cannot be reached solely from evidence in another. Third, it predicts differential vulnerability. If two technologies are equally identity-mediating but differ in *C*, interruption should have different cognitive consequences even if both threaten self-presentation. Fourth, it derives different safeguards through foreseeable dependence and asymmetric control. A classification can therefore be mistaken: evidence may fail to support the assigned category, predicted vulnerabilities may not occur, or the proposed safeguard may not address the relevant interest. These possibilities make the framework more than an unfalsifiable vocabulary.

The framework remains intentionally pluralist. Domain-specific theories still determine whether cognition is genuinely extended, what constitutes personhood, and which capacities ground responsibility. The proposal’s distinctive role is to coordinate those theories without allowing their evidential standards to migrate unexamined across domains.

7.2 Objection: integration cannot answer the coupling–constitution problem

A critic of extended cognition may accept all the proposed indicators but insist that they show only unusually intimate causal coupling. Reliability, personalization, and trust do not transform an external cause into a constituent of cognition.

The framework is compatible with disagreement at the constitutive threshold. Its graded concept of integration is not intended to define the controversy away. Instead, it separates two questions: how tightly an artifact is organized into cognitive activity, and whether sufficiently tight organization warrants a constitutive description. An intracranialist and an extended-mind theorist can agree that one relation is more integrated than another while disagreeing about the metaphysical conclusion. This common ground is valuable for evaluating digital cases.

Moreover, the framework resists the inflation that motivates the objection. It does not treat access to the internet or a single model interaction as cognitive extension. A constitutive claim becomes plausible only when reciprocal dependence, stability, availability, personalization, and transformation converge. Even then, the claim is about a specific process, not the wholesale location of a person's mind. The burden of argument remains visible.

7.3 Objection: digital selves are only representations

Profiles and avatars, the objection continues, may influence a person, but representations of a thing are not parts of the thing. A portrait is not part of the sitter, and a profile is not part of its user.

The reply is that the self-representational dimension concerns narrative and practical selfhood, not the material composition of a biological organism. Representations can be constitutive within a practice when the practice depends on self-interpretation and recognition. A promise is expressed through signs, but this does not make its role in a commitment unreal. Similarly, a publicly sustained profile can organize relationships, preserve autobiographical materials, and guide future action. Its constitutive relevance derives from these relations, not from pictorial resemblance.

The framework also leaves room for degrees and failure. Most representations remain peripheral. Some become central but alienating. Stronger realist accounts of digital selves, such as that of Declos and Grandjean (2026), may argue that certain digital entities are genuine components of the self. The present proposal accepts that possibility at the level of selfhood while refusing an automatic inference to numerical identity or personhood. Thus it captures the practical force of the digital-self claim without requiring every avatar to be literally identical with its user.

7.4 Objection: hybrid agency anthropomorphizes artificial intelligence

To call an artificial system a participant in agency may project intentions and understanding onto statistical machinery. The terminology could encourage users to trust fluent systems or allow institutions to shift blame to them.

This concern supports, rather than defeats, the framework's distinction. Agential distribution is analyzed through functions—goal formation, option generation, selection, execution, monitoring—not through surface resemblance to conversation. The account does not infer inner experience from output. A system may occupy some functional roles while lacking autonomous goals, self-maintenance, understanding, or moral reasons. Describing this limited role is less anthropomorphic than calling the system either a person or a neutral conduit.

The vocabulary also helps prevent responsibility laundering. If an organization knows that an artificial component significantly structures options or executes actions, it cannot plausibly describe harmful outputs as unforeseeable events produced by an independent electronic person. The correct question becomes who designed and governed the distribution of control. Thin functional agency is compatible with strong human and institutional accountability.

7.5 Objection: phenomenal consciousness should be a fourth dimension

Embodied and enactive theories challenge the idea that consciousness must be brain-bound. If experience depends on active engagement with the environment, a virtual environment or technological interface might partly constitute it. Excluding phenomenal realization from the profile may therefore appear to reproduce internalism or make the framework incomplete.

The exclusion is methodological rather than metaphysical. C, S, and A diagnose relations that can be investigated through recurrent use, autobiographical and social practices, and the allocation of control. Phenomenal consciousness is not another degree of the same kind. Its attribution depends on a theory of consciousness and on evidence that the relevant realizing conditions are present. Treating it as a parallel ordinal score would conceal rather than solve that theory dependence.

Nothing in the framework rules out extended or artificial consciousness. An enactive theory may identify sensorimotor or organizational relations sufficient for experience, and a future architecture may satisfy well-supported indicators. In such a study, phenomenal realization can be represented as a separate theory-indexed assessment, $P(T)$, where T identifies the theory supplying the criteria. It should not be inferred from C, S, or A. This preserves openness to extended consciousness while preventing fluency, immersion, or practical importance from functioning as surrogate evidence of subjectivity.

8 Implications for Epistemology, Ethics, and Design

The framework does not derive ethical duties from classification alone. A mediated profile identifies the interest placed at risk and the mechanism of vulnerability. A further normative argument identifies responsible actors by asking who invited or benefited from reliance, who controls the relevant infrastructure, who could reasonably prevent harm, and who occupies a professional or institutional role. This two-step procedure prevents both technological determinism and a generic list of “responsible AI” values.

8.1 Cognitive integration: epistemic and continuity obligations

Trust should be calibrated to the role a system occupies in cognition. As C rises, use becomes practically transparent: the person consults the resource with less deliberation because it has become part of a familiar routine. Practical transparency can coexist with evidential opacity. Conversational fluency is not provenance, personalization is not reliability, and habitual access is not a guarantee of continuity.

Systems intended for high-C use should therefore distinguish source material from model inference, preserve correction histories, support verification, and permit export into functionally usable formats. Continuity measures are not merely conveniences when interruption would impair an acquired capacity. Providers that encourage users to externalize memory or planning while retaining unilateral power to alter access create a foreseeable dependency and acquire a correspondingly stronger duty to provide notice, recovery, and portability.

8.2 Self-representational mediation: contestability and contextual integrity

Digital selfhood is negotiated among persons, audiences, platforms, communities, and automated classifications. As S rises, ethical evaluation should ask who can revise a representation, which past versions persist, whether context travels with archived material, and when an inferred trait is treated as authoritative. The relevant protection is not an unrestricted right to rewrite every social record. It is a stronger claim to inspect automated inferences, correct factual error, append context, contest imposed categorization, and prevent impersonation when a system has acquired autobiographical or recognitional authority.

The relational perspective matters here. A profile can affect not only private self-esteem but also family roles, professional standing, communal membership, and access to institutional opportunities. The same platform feature may therefore have different identity significance across cultural and social settings. High S must be assessed through the importance of the mediated relation, not through a universal assumption that public visibility or expressive freedom has the same meaning for every user.

8.3 Agential distribution: meaningful control and responsibility mapping

“Human in the loop” is inadequate when the human merely approves options that an opaque system has already framed and cannot competently evaluate. Oversight should be matched to the actual allocation of goal formation, option generation, selection, execution, and monitoring. Following Santoni de Sio and van den Hoven (2018), meaningful control requires both tracking and tracing: system behavior must remain responsive to relevant human reasons, and outcomes must be traceable to agents who understand their role and can be called to account.

For high-A systems, this entails an ability to interrupt or reverse delegated action, logs that reconstruct significant selections, calibrated explanation of the system's discretionary role, and an advance responsibility map covering design, deployment, authorization, monitoring, and remediation. A nominal final click does not restore control when earlier stages have already narrowed the decision space. Conversely, automated execution does not eliminate human agency when responsible actors set goals, understand constraints, monitor performance, retain effective stop conditions, and can remedy foreseeable failure.

The framework thus yields a focused research agenda. Empirical work can test whether the diagnostic criteria reliably distinguish patterns of dependence, recognition, and control across users and cultures. Normative analysis can refine the thresholds at which provider obligations intensify. Philosophy of mind can investigate whether particular high-bandwidth relations satisfy a theory-indexed assessment of phenomenal realization. These projects are connected, but none can be answered by the single metaphor of a "digital consciousness."

9 Conclusion

Digital technologies transform human life, but they do not do so through one metaphysical operation. The language of a blurred boundary between consciousness and technology conceals at least three different changes. Artifacts can become integrated into cognitive processes. Digital representations can participate in narrative and practical selfhood. Artificial systems can redistribute the production and control of action. These changes can occur together, but none entails the others.

The proposed framework represents a mediated configuration as a profile of cognitive integration, self-representational mediation, and agential distribution. Its principal philosophical value lies in preserving the boundaries of inference. Cognitive extension does not by itself establish extended phenomenal consciousness. Multiple profiles and avatars do not by themselves create multiple persons. Hybrid agency does not by itself confer moral responsibility or personhood on artificial intelligence. Each stronger conclusion requires an additional theory and independent evidence.

This disciplined account is not technologically conservative. It recognizes that a non-conscious artifact can become deeply important to memory, identity, and action. It also reveals dimensions of vulnerability that an exclusive focus on machine consciousness would miss. The loss of an integrated memory system, the manipulation of an identity-bearing profile, and the opaque delegation of action are different harms with different remedies.

The most defensible conclusion is therefore relational. Digital technologies alter the arrangements through which persons think, understand themselves, and act. Their philosophical significance lies not in the emergence of a unitary "digital consciousness," but in the changing architecture of human cognitive dependence, self-interpretation, and control.

Declarations

Use of Generative AI

During the preparation of this work, OpenAI ChatGPT was used to assist with literature organization, argument restructuring, English-language drafting, and language editing. All scholarly judgments, source verification, and responsibility for the submitted content remain with the human author.

Funding

No funding was received for conducting this study or preparing this manuscript.

Author Contribution

S.A. conceptualized the study, conducted the literature analysis, developed the three-dimensional framework, and wrote and revised the manuscript. S.A. approved the final version.

References

1. Adams F, Aizawa K (2001) The bounds of cognition. *Philosophical Psychol* 14(1):43–64. <https://doi.org/10.1080/09515080120033571>
2. Ames RT (2011) *Confucian Role Ethics: A Vocabulary*. University of Hawai'i, Honolulu
3. Aydin C, González Woge M, Verbeek PP (2019) Technological environmentality: Conceptualizing technology as a mediating milieu. *Philos Technol* 32:321–338. <https://doi.org/10.1007/s13347-018-0309-3>
4. Barandiaran XE, Almendros LS (2025) Transforming agency: On the mode of existence of large language models. *Phenomenology Cogn Sci*. <https://doi.org/10.1007/s11097-025-10094-3>
5. Block N (1995) On a confusion about a function of consciousness. *Behav Brain Sci* 18(2):227–247. <https://doi.org/10.1017/S0140525X00038188>
6. Bortolan A (2024) Becoming oneself online: Narrative self-constitution and the internet. *Philos Stud* 181:2405–2427. <https://doi.org/10.1007/s11098-024-02169-9>
7. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A et al (2023) Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. <https://doi.org/10.48550/arXiv.2308.08708>
8. Chalmers DJ (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York
9. Chalmers DJ (2023) Could a large language model be conscious? *arXiv*. <https://doi.org/10.48550/arXiv.2303.07103>

10. Clark A (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, New York. <https://doi.org/10.1093/acprof:oso/9780195333213.001.0001>
11. Clark A (2009) Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind* 118(472):963–993. <https://doi.org/10.1093/mind/fzp110>
12. Clark A, Chalmers DJ (1998) The extended mind. *Analysis* 58(1):7–19. <https://doi.org/10.1093/analys/58.1.7>
13. Declos A, Grandjean V (2026) Digital selves. *Synthese* 208:42. <https://doi.org/10.1007/s11229-026-05705-8>
14. Floridi L (2014) *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press, Oxford
15. Floridi L (2023) AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philos Technol* 36:15. <https://doi.org/10.1007/s13347-023-00621-y>
16. Floridi L, Sanders JW (2004) On the morality of artificial agents. *Mind Mach* 14:349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
17. Heersmink R (2013) Embodied tools, cognitive tools and brain–computer interfaces. *Neuroethics* 6:207–219. <https://doi.org/10.1007/s12152-011-9136-2>
18. Heersmink R (2015) Dimensions of integration in embedded and extended cognitive systems. *Phenomenology Cogn Sci* 14(3):577–598. <https://doi.org/10.1007/s11097-014-9355-1>
19. 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>
20. Heersmink R (2018) The narrative self, distributed memory, and evocative objects. *Philos Stud* 175(8):1829–1849. <https://doi.org/10.1007/s11098-017-0935-0>
21. Heersmink R (2020) Narrative niche construction: Memory ecologies and distributed narrative identities. *Biology Philos* 35:53. <https://doi.org/10.1007/s10539-020-09770-2>
22. Heersmink R, de Rooij B, Clavel Vázquez MJ, Colombo M (2024) A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics Inf Technol* 26:41. <https://doi.org/10.1007/s10676-024-09777-3>
23. Kudina O, Ballsun-Stanton B, Alfano M (2025) The use of large language models as scaffolds for proleptic reasoning. *Asian J Philos* 4:24. <https://doi.org/10.1007/s44204-025-00247-1>
24. Menary R (ed) (2010) *The Extended Mind*. MIT Press, Cambridge, MA
25. Menkiti IA (1984) Person and community in African traditional thought. In: Wright RA (ed) *African Philosophy: An Introduction*, 3rd edn. University Press of America, Lanham, MD, pp 171–181
26. Nyholm S (2018) Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Sci Eng Ethics* 24:1201–1219. <https://doi.org/10.1007/s11948-017-9943-x>
27. Parfit D (1984) *Reasons and Persons*. Clarendon, Oxford

28. de Santoni F, van den Hoven J (2018) Meaningful human control over autonomous systems: A philosophical account. *Front Rob AI* 5:15. <https://doi.org/10.3389/frobt.2018.00015>
29. Schechtman M (1996) *The Constitution of Selves*. Cornell University Press, Ithaca, NY
30. Schechtman M (2014) *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford University Press, Oxford
31. Smart PR (2017) Extended cognition and the Internet. *Philos Technol* 30:357–390. <https://doi.org/10.1007/s13347-016-0250-2>
32. Smart PR, Clowes RW, Clark A (2025) ChatGPT, extended: Large language models and the extended mind. *Synthese* 205:242. <https://doi.org/10.1007/s11229-025-05046-y>
33. Smart PR, Clowes RW, Krueger J, Boniface M (2026) The story of your life: Large language models and personal memory. *Rev Philos Psychol*. <https://doi.org/10.1007/s13164-026-00831-1>
34. Turkle S (1995) *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York
35. Véliz C (2021) Moral zombies: Why algorithms are not moral agents. *AI Soc* 36:487–497. <https://doi.org/10.1007/s00146-021-01189-x>
36. Verbeek PP (2005) *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Pennsylvania State University Press, University Park, PA
37. Verbeek PP (2008) Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology Cogn Sci* 7(3):387–395. <https://doi.org/10.1007/s11097-008-9099-x>