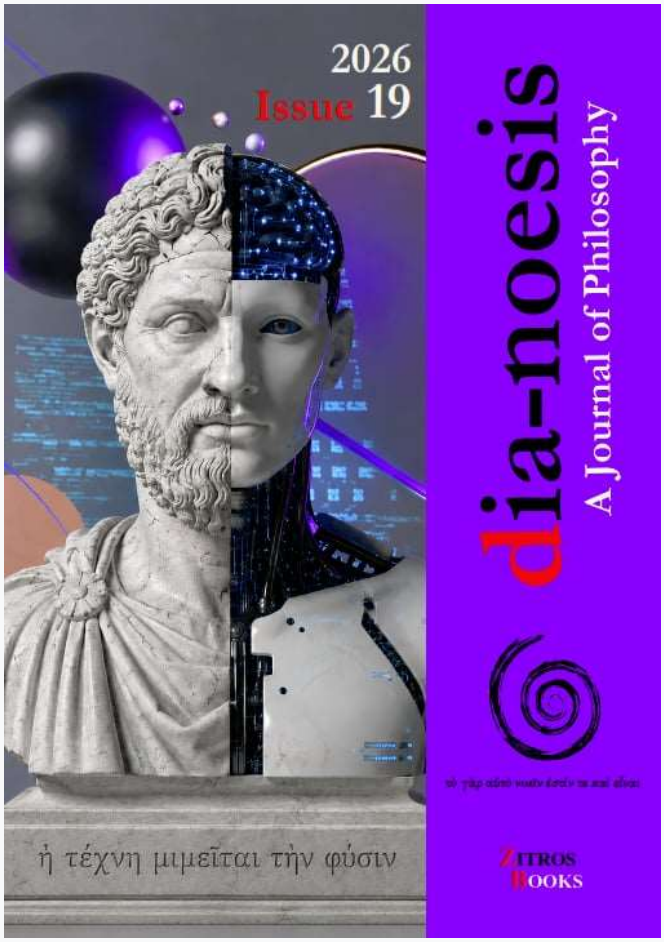
Dia-noesis: A Journal of Philosophy

Vol 19, No 1 (2026)

Philosophy and Artificial Intelligence: Rethinking the Human and the Political



The Erosion of Subjectivity

Lesia Hrechukha, Olena Pavlenko, Natalia Butko

doi: 10.12681/dia.45793

To cite this article:

Hrechukha, L., Pavlenko, O., & Butko, N. (2026). The Erosion of Subjectivity: Cognitive Offloading and the Responsibility Gap in the Age of Algorithmic Governance. *Dia-Noesis: A Journal of Philosophy*, *19*(1), 137–158. https://doi.org/10.12681/dia.45793

The Erosion of Subjectivity: Cognitive Offloading and the Responsibility Gap in the Age of Algorithmic Governance

Lesia Hrechukha,
Ph.D., Associate Professor,
Cherkasy State Technological University
l.hrechukha@gmail.com

Olena Pavlenko,
Ph.D., Professor,
Mariupol State University,
h.pavlenko@mu.edu.ua

Nataliia Butko,
Ph.D., Associate Professor,
Bohdan Khmelnytsky
National University of Cherkasy,
butko_n_v@ukr.net

Abstract[1]

This study examines the ontological and ethical transformations brought about by the integration of advanced AI into social, educational, and communicative domains. Adopting a relational rather than instrumentalist perspective, it

[1] The authors declare that Gemini AI was utilized during the final stage of this manuscript. Specifically, the tool was used to refine linguistic fluency, and organize bibliographic data due to the requirements. The authors maintain full responsibility for the final content and intellectual integrity of the work.

investigates how algorithmic agency blurs the boundaries of human cognition and subjectivity. Through conceptual engineering, the study identifies deficiencies in folk conceptions of responsibility and proposes a shift toward prospective accountability. A central finding is the "phenomenological trap" generated by the functional simulation of agency in Large Language Models, which induces moral deskilling and an illusion of control — a claim corroborated by psychophysiological (EEG) data revealing a neurological decoupling between implicit agency and externalized moral responsibility. The study further explores cognitive offloading and the emergence of epistemic dependence, and critically assesses current regulatory frameworks such as the EU AI Act, arguing for a "friction-by-design" approach. It concludes that preserving human identity in AI-mediated societies requires the conscious design of sociotechnical systems that supplement rather than replace human interpretive judgment.

Keywords: algorithmic agency, cognitive offloading, digital colonialism, epistemic dependence, friction-by-design, moral deskilling, responsibility gap, prospective accountability.

Introduction

The pervasive integration of Large Language Models (LLMs) and autonomous agents into the fabric of daily intellectual labor marks a definitive "relational turn" in the philosophy of technology. No longer confined to specialized industrial tasks, AI now mediates the most fundamental aspects of human cognition, from educational assessment to moral deliberation. This shift renders the classical instrumentalist view where technology is considered a neutral, passive tool entirely dependent on human intent[2]. As contemporary systems exhibit what Luciano Floridi[3] calls "agency without intelligence," they cease to be simple instruments and instead become active co-participants in the formation of epistemic agency, influencing how we form, revise, and justify our beliefs.

The emergence of systems capable of autonomous processing creates a fundamental philosophical challenge that forces a reassessment of core concepts of reason and freedom.[4] However, current scholarly orientations often fail to reconcile this functional agency with traditional metaphysical categories of responsibility. While the "responsibility gap" was historically framed as a

[2] Verbeek, 2005.
[3] Floridi, 2023.
[4] Papacharalambous, 2021.

technical byproduct of complex systems[5], the delegating of critical decisions to invulnerable algorithms now creates an ethical vacuum. Within this normative vacuum, the algorithm fails to qualify as a moral subject. Furthermore, the human agent systematically relinquishes effective control over the process, leading to a state of moral deskilling that is a gradual atrophy of the capacity for autonomous ethical reasoning[6].

The necessity to rethink accountability is further driven by the transition to "AI-mediated societies," where algorithms act as primary instruments of institutional and political governance. This environment of "algocracy" introduces specific forms of vulnerability and novel power structures. [7] When algorithmic decisions become routine in education and communication, citizens structurally turn into "moral patients"[8]. In this very case, they function as subjects of a system that lacks moral reciprocity. This structural asymmetry, combined with technical opacity, threatens the very possibility of existing as self-determining subjects in a democratic society.

This study aims to examine the foundational ontological categories called into question by these systems: what it means to possess agency, where the boundaries of human cognition lie, and how automation transforms human subjectivity. To achieve this, the article analyzes the ontological status of AI agents, evaluates the psychological impact of cognitive offloading based on recent empirical data, and explores the systemic implications for educational and communicative ethics. Without a rigorous philosophical intervention, we risk losing not only control over our technologies but the very capacity for autonomous choice that defines human identity.

Methodology

To address the ethical vacuums created by autonomous systems, the study employs the method of conceptual engineering, as

[5] Matthias, 2004: 175–183.
[6] Vallor & Vierkant, 2024.
[7] Danaher, 2016: 245–268; cf. Vavouras et al., 2024.
[8] Floridi, 2014.

articulated in the context of AI ethics by Herman Veluwenkamp[9]. This normative approach proceeds from the realization that our traditional, "folk" concept of responsibility is no longer "fit for purpose" in an age dominated by algorithmic agency. Unlike descriptive analysis, which merely seeks to define how we currently use a term, conceptual engineering is a prescriptive process. It aims to improve our representational tools to serve better our cognitive and ethical goals.

The necessity of this method stems from the inherent "defects" in the classical metaphysical category of responsibility when applied to sociotechnical networks. Traditionally, responsibility requires two conditions. Firstly is to control. It means the agent must have causal power. Secondly, it is knowledge, as the agent has to understand the consequences. However, as AI systems introduce epistemic opacity (the "black box" problem) and autonomous subgoal formation, these conditions are systematically undermined. We cannot simply use the "old" word responsibility because it was designed for human-to-human interactions characterized by mutual vulnerability and transparency.

The application of conceptual engineering in this research follows a rigorous three-stage process. The initial stage is a diagnostic one, so called conceptual evaluation. We first analyze the traditional retributive and merit-based models of responsibility. This stage identifies the "defects" of these concepts when faced with the "responsibility gap"[10]. Specifically, how they fail to assign blame when an algorithm's action is neither strictly deterministic nor truly intentional. The next stage is to identify conceptual gaps. Due to Veluwenkamp's engineering of the term, we distinguish between epistemic responsibility gaps (erg), where opacity prevents identification of the agent, and control misalignments, where the distribution of control is suboptimal for harm minimization. And the final stage is an ameliorative one or a conceptual redesign. Finally, we might offer a "re-engineered" version of responsibility, namely a prospective accountability. This new structure shifts the focus from retrospective blame (finding a single guilty party after the fact) to a forward-looking, relational practice. This engineered concept prioritizes "ethics by design" and collective governance,

[9] Veluwenkamp, 2025.
[10] Matthias, 2004: 175–183.

making accountability a precondition for the deployment of hybrid systems rather than a reaction to their failure.

By utilizing conceptual engineering, this study does not merely describe the ethical crisis of AI but actively reconfigures our moral vocabulary to preserve human agency in an algocratic reality. This methodological rigor ensures that our findings are not only philosophically sound but also practically applicable to the legislative and ethical regulation of emerging technologies.

Findings

A primary objective of recent philosophical inquiry has been to definitively resolve the ontological status of artificial agency, moving beyond outdated anthropomorphic comparisons. Recent scientific works systematically distinguish between functional (or imitative) agency and genuine moral agency. Formosa P., Hipólito I. and Montefiore T.[11] in their comprehensive analysis of human-machine interaction demonstrate that contemporary generative AI exhibits only a highly sophisticated simulation of agency. By evaluating large language models against the rigid philosophical criteria of basic, autonomous, and moral agency, the authors conclude that these systems remain entirely bound by pre-programmed objectives and statistical training parameters. Due to a lack of phenomenal consciousness and authentic goal-formation, their outputs constitute mechanical pattern matching rather than true, deliberative intent. This modern empirical reality forcefully corroborates John Searle's foundational "Chinese Room"[12] argument that tells that contemporary AI produces a compelling illusion of semantic understanding without possessing any genuine internal cognitive states.

However, the academic discourse does not end at pure dismissal. Recognizing the practical reality of how humans interact with advanced systems, Martela[13] underlines the vital nuance of functional agency. The scientist examines the LLM-driven autonomous "Voyager" agent within the digital ecosystem of Minecraft, arguing that applying Daniel Dennett's "intentional stance" is practically

[11] Formosa et al., 2025.
[12] Searle, 1980: 417–457.
[13] Martela, 2025.

necessary. As systems like Voyager can autonomously formulate sub-goals and dynamically alter behavior based on real-time sensory feedback, their actions are accurately predicted by ascribing them intentionality. Bridging these two perspectives is the relational turn in the philosophy of technology, Czyszczoń M.[14] introduces the RI_3 matrix, a framework where moral agency is not an intrinsic property but emerges organically from sociotechnical networks where action and responsibility are dynamically distributed among human actors and machine artefacts.

The ontological ambiguity of artificial agents is not merely a byproduct of their computational complexity, but is fundamentally rooted in the human phenomenological tendency to project interiority onto sophisticated behavioral patterns. Despite the established ontological premise that artificial intelligence lacks phenomenal consciousness, the lived experience of the user frequently contradicts this metaphysical reality[15]. This discrepancy necessitates a deeper investigation into the psychological mechanisms of anthropomorphism and the cognitive "traps" inherent in human-AI interaction.

The key role in this discussion is played by Daniel Dennett's "Intentional Stance"[16]. While Martela F.[17] utilizes this concept as a pragmatic predictive tool for autonomous agents like Voyager, a more critical reading reveals its role as a psychological catalyst for ontological confusion. Dennett posits that when we encounter a system of sufficient complexity, we find it cognitively "cheaper" to treat it as an agent with beliefs and desires rather than analyzing its physical or design-based constraints. However, in the age of Large Language Models, this stance has evolved from a voluntary predictive strategy into an involuntary phenomenological trap. The linguistic fluency of LLMs triggers deeply ingrained social heuristics, forcing the human brain to process the machine as a "communicative subject" rather than a statistical processor.

As argued by Sven Nyholm in "Humans and Robots: Ethics, Agency, and Anthropomorphism"[18], our tendency to anthropomorphize is not a simple intellectual error, but a structural feature

[14] Czyszczyń, 2025.
[15] Formosa et al., 2025.
[16] Dennett, 1987.
[17] Martela, 2025.
[18] Nyholm, 2020.

of human social cognition. Nyholm emphasizes that when robots or AI systems mimic human-like reactive behaviors, for instance, empathy, hesitation, or humor, they "hook" into our evolutionary predisposition for social engagement. This creates a state of "Value-Sensitive Design" in reverse, where the interface design actively encourages the user to overlook the "responsibility gap" by cloaking the algorithm in the familiar garb of personhood. Nyholm warns that this creates a dangerous "asymmetry of perceived agency" where the user subjectively feels they are interacting with a peer, which leads to a premature transfer of trust and a subsequent erosion of the user's own critical distance.

Furthermore, the phenomenological experience of AI interaction is characterized by what might be termed "semantic seduction". As LLMs operate within the medium of natural language, that is the primary vehicle for human consciousness, the user experiences the output not as a calculated probability but as an intentional speech act. This leads to the "intentional trap", even when a user intellectually understands that the system is a "black box" without internal states[19]. The experiential reality of the dialogue fosters an illusion of shared understanding. This illusion is the cornerstone of moral deskilling[20]. The human agents abdicate their role as the sole source of meaning and judgment by treating the algorithm as an interlocutor with its own perspective.

Consequently, the ontological status of AI must be understood as a relational construct. As Nyholm S.[21] suggests, the machine does not need to be a person to function as one within our moral ecology. The danger lies in the "ontological slippage" where the functional simulation of intent becomes a surrogate for genuine moral reciprocity. By expanding our investigation into these phenomenological dimensions, it becomes clear that the "responsibility gap" is not only a technical or legal failure, but a psychological one, sustained by our own cognitive tendency to find "someone" where there is only "something".

The seamless integration of generative AI into daily intellectual labor has precipitated a radical reconceptualization of the boundaries of human cognition. Applying the theory of the extended

[19] Pasquale, 2015.
[20] Vallor & Vierkant, 2024.
[21] Nyholm, 2020.

mind, Clark A.[22] argues that humans act as "natural-born cyborgs" whose cognitive processes extend into their digital environment. From this optimistic perspective, generative AI acts as a monumental expansion of the digital unconscious, offering syntactic structures and ideas that the user incorporates as an extension of their own thought process. When paired with rigorous "epistemic hygiene," AI becomes an invaluable resource for active inference and resolving environmental uncertainty.

However, this theoretical optimism is undercut by empirical data. Gerlich M.[23] conducted a comprehensive study involving 666 participants, establishing a highly significant positive correlation between the frequency of AI use and cognitive offloading, alongside an alarming negative correlation with critical thinking capacities. The study highlights that younger demographics exhibit significantly greater structural dependency, scoring lowest on independent analytical metrics. When automation is applied to research, it induces the premature delegation of analysis, drastically diminishing internal cognitive engagement. Consequently, the boundaries of cognition are displaced inward, marking an epistemic retreat. Human thought effectively terminates the moment a user passively accepts an algorithm's output without verification. This poses the risk of the "hollowed mind" in which an intellect processes atrophy through its disuse.

A critical dimension of the automation of intellectual labor is the profound discrepancy between the neurological feeling of control and the objective exercise of responsibility. This study argues that contemporary AI interfaces are often designed to mask the delegation of judgment, creating what cognitive scientists call the "illusion of control". This phenomenon is not merely a psychological byproduct but a fundamental transformation of human subjectivity within sociotechnical networks.

To understand this shift, we have to examine the psychophysiological data from Salatino A., Prével, A et al.[24] in their "drone cadet" study. Using electroencephalography (EEG)[25], researchers

[22] Clark, 2025.

[23] Gerlich, 2025.

[24] Salatino et al., 2025.

[25] EEG, as a high-resolution temporal neuroimaging tool, allowed researchers to detect Event-Related Potentials (ERPs) that confirm the brain's implicit 'reward' for automated actions, even when explicit judgment is absent.

tracked the "intentional binding" effect. It is a neurological marker where the perceived time between an action and its effect is shortened, indicating a strong internal sense of agency. The data revealed that even when the AI algorithm had already pre-selected the target, participants' brains continued to register a high implicit sense of agency. However, when errors occurred, this feeling did not translate into an explicit assumption of responsibility. Participants reflexively externalized the blame to the "black-box" system. This psychophysiological decoupling suggests automation "hacks" the human brain's agency circuits, creating a state of moral anesthesia where the operator remains neurologically engaged but ethically detached.

The inward displacement of cognitive boundaries, as identified in the preceding analysis of offloading[26], precipitates a fundamental shift in the human epistemic condition. In addition, it might lead to the emergence of a profound and systemic epistemic dependence. This study argues that as Large Language Models transition from mere "search assistants" to "autonomous synthesizers" of knowledge, the human agent is no longer merely assisted by technology but becomes structurally reliant upon it for the very validation of truth. This reliance generates what we term a "verification crisis" that is a state where the convenience of algorithmic output systematically undermines the user's capacity (and willingness) to engage in the frictional labor of cross-verification.

The gradual erosion of first-order expertise characterizes epistemic dependence in the era of algorithmic governance. In professional domains (ranging from applied linguistics and translation to medical diagnostics) expertise is traditionally cultivated through a "frictional" engagement with complex, often contradictory data. This cognitive friction is the essential catalyst for the development of phronesis (practical wisdom). However, when AI mediators provide syntactically fluent and seemingly authoritative summaries, they remove the necessity for this engagement. As a result, the expert gradually becomes a "passive monitor" of a black-box system. The risk here is not just the occurrence of "hallucinations" or algorithmic errors, but the atrophy of the human capacity to detect them. If a human agent lacks the domain-specific depth to challenge the machine, the "responsibility gap" expands from a

[26] Gerlich, 2025.

technical void to an existential one. Then we become responsible for decisions whose internal logic we can no longer justify or even comprehend.

In addition, this dependence fosters the creation of "epistemic bubbles of automation". Unlike traditional filter bubbles, which limit the scope of information, automation bubbles limit the depth of interpretation. When users accept an LLM's synthesis, they are not just receiving data. Contrary, they are adopting a pre-packaged, homogenized worldview that reflects the statistical biases of the training set rather than the nuanced reality of the specific case. This leads to semantic erosion, which is a process where the richness of professional language and the precision of cultural context are sacrificed for the sake of frictionless efficiency. In the context of AI-mediated societies, this dependency has a systemic political dimension. When institutions delegate critical judgment to algorithms, they create a "black-box society"[27] where truth is no longer a matter of transparent deliberation but a byproduct of proprietary computation. This structural opacity renders democratic oversight impossible, as the "experts" tasked with regulating the technology are often themselves epistemically dependent on the very systems they seek to govern.

Finally, the transition toward epistemic dependence marks the final stage of moral deskilling[28]. By abdicating the right to judgment and the duty of verification, the subject forfeits their role as an autonomous epistemic agent. To preserve human identity in this landscape, it is imperative to design "friction-heavy" interfaces that actively demand human intervention and cross-verification, thereby preventing the transformation of the expanded mind into a hollowed intellect.

In the realm of intercultural communication, the integration of advanced Neural Machine Translation (NMT) has introduced a profound paradox. While global discourse is vastly accelerated, the depth of semantic exchange is systematically compromised. This study argues that AI, operating as an invisible mediator, executes translation tasks through a process of mechanical pattern matching that inherently disregards the cultural lived experience of the interlocutors. Translation, in its authentic form, is not merely the

[27] Pasquale, 2015.
[28] Vallor & Vierkant, 2024.

replacement of syntactic units. It is a "transfer of meanings" that requires a relational awareness and a degree of human vulnerability that algorithms fundamentally lack.

A critical consequence of this automated mediation is the emergence of "semantic homogenization". As algorithms rely on vast datasets that favor standardized linguistic structures, they tend to smooth over cultural idiosyncrasies, local metaphors, and sociolinguistic nuances. In diplomatic and literary contexts, this results in a literal, "smooth" text that lacks the underlying cultural subtext. For instance, in diplomatic discourse, where the precision of idiomatic expression is vital for conflict resolution, the uncritical use of AI can lead to catastrophic misinterpretations. Furthermore, the automation of translation induces a structural risk of semantic simplification in specialized domains such as law. Legal terms are not universal constants but they are deeply embedded in specific historical frameworks. When an algorithm translates these terms based on statistical frequency, it strips the language of its historical weight. The translator is then reduced from an active cultural author to a mere "ethical buffer". This process might contribute to moral deskilling[29], where the user ceases to perceive ethical problems in communication simply because the algorithm has produced a syntactically fluent but culturally hollow text.

The integration of AI into educational oversight, via automated assessment and biometric proctoring, engenders a profound crisis of pedagogical accountability. AI-driven proctoring systems, analyzing facial expressions and physical presence, transfer the right to adjudicate learning outcomes directly to algorithms. However, these systems often reproduce structural biases, routinely flagging students with darker skin tones or neurodivergent behaviors as "suspicious" based on biased historical data.

This delegation of judgment triggers specific ethical vacuums, identified by Veluwenkamp H.[30] as Epistemic Responsibility Gaps (ERG) and Control Misalignments (CM). An ERG occurs when algorithmic opacity impedes the identification of the responsible party for an unjust penalty. It is unclear if the fault lies with the instructor's prompt, the university's policy, or the developer's neural network. A Control Misalignment happens when the

[29] Ibid.
[30] Veluwenkamp, 2025.

responsible party, that is an instructor, is known, but they lack the technical capacity to override automated harm. Beyond legal liability, excessive automation fundamentally alters identity formation. The student becomes a passive user navigating algorithmic pathways, effectively transforming into an "optimized data profile". This environment minimizes academic friction and isolates the student, resulting in severe epistemic deskilling and the fragmentation of professional identity.

On a macro-political scale, the transition toward AI-mediated societies necessitates a fundamental re-evaluation of structural power. The reliance on algocracy[31] threatens academic freedom and democratic agency through opaque digital surveillance. Algorithms covertly impose methodological standards through predictive analytics, effectively bypassing democratic deliberation. Traditional models of retrospective legal liability prove ineffectual for hybrid intelligent systems.

As Nyholm S.[32] emphasizes, there is an absolute necessity to transition toward prospective accountability through "ethics by design". Without this, users become mere "patients" of systems that lack moral reciprocity[33]. As UNESCO[34] and scholars of digital ethics consistently warn, the uncritical deployment of these tools intensifies the risk of algorithmic discrimination and digital colonialism. Global inequality is intensified, as power is concentrated in a few transnational technology corporations, creating a monochromatic digital culture that prioritizes corporate predictability over cultural authenticity and democratic participation.

The philosophical identification of the "responsibility gap"[35] and the psychophysiological "illusion of control"[36] demand a radical reassessment of current legislative efforts to govern artificial intelligence. While the global regulatory landscape has been dominated by the phased implementation of the EU AI Act (Regulation (EU) 2024/1689), this study argues that the Act's primary architecture, which is centered on "risk-based categorization", remains ontologically insufficient. By classifying AI systems into categories of

[31] Danaher, 2016: 245–268.
[32] Nyholm, 2020.
[33] Floridi, 2014.
[34] UNESCO, 2025.
[35] Matthias, 2004: 175–183.
[36] Salatino et al., 2025.

"unacceptable," "high," "limited," and "minimal" risk, the European legislator attempts to treat algorithmic agency as a manageable industrial hazard. However, this taxonomic approach fails to address the underlying erosion of human subjectivity that occurs even within "limited" risk interactions, such as those mediated by LLMs in intellectual labor. To preserve human identity in AI-mediated societies, regulation must transition from a reactive model of retrospective liability to a robust, forward-looking framework of prospective accountability.

A core deficiency in the current legal framework is the excessive reliance on the "Human-in-the-loop" (HITL) requirement as a panacea for algorithmic opacity. Article 14 of the AI Act mandates that high-risk natural person's design systems in a way that allows for effective oversight. Yet, our empirical findings suggest that HITL is often a cognitive fallacy. Because of the "intentional binding" effect and the subsequent moral anesthesia[37], a human operator often acts as a pure "rubber stamp" for automated decisions. When an algorithm operates at speeds or levels of complexity that exceed human cognitive processing, the "loop" becomes a formalistic ritual rather than a substantive check.

To counter this, we offer a regulatory shift toward "friction-by-design". This principle codified into the technical standards of the next generation of AI regulations would legally mandate that high-risk systems, particularly in education, translation, and law, include "mandatory cognitive checkpoints." Unlike current seamless interfaces that prioritize "frictionless" output, these checkpoints would force intervals of human deliberation. For instance, an AI-mediated legal or pedagogical assessment system would require the human supervisor to reconstruct and justify the algorithm's logic in natural language before the decision is finalized. This move from a passive "loop" to an active "Human-on-the-loop" (HOTL) oversight is essential for preventing the epistemic atrophy of professional expertise. Without legally mandated "cognitive friction," the human agent remains a prisoner of the interface, assuming responsibility for outcomes they did not truly deliberate upon.

To address the tension between user-centric interface design and normative ethics, it is necessary to establish a distinction between "seamlessness" as a commercial imperative and "ethical friction" as

[37] Ibid.

a mechanism for preserving human subjecthood. Within the digital economy[38], seamlessness is engineered to eliminate any barrier to continuous user engagement, prioritizing frictionless efficiency over deliberate choice[39]. Psychologically, this hyper-optimized architecture systematically exploits what Daniel Kahneman[40] identifies as "System 1" cognition - the fast, automatic, and highly intuitive mode of thought that requires minimal effort but is fundamentally vulnerable to anthropomorphic bias and the "illusion of control." By confining the user in a state of frictionless automation, the interface actively induces uncritical acceptance and moral deskilling. To counter this, "friction-by-design" must be implemented not as a regression in usability, but as a deliberate sociotechnical mechanism intended to jolt the user into "System 2" processing. This slower, analytical, and effortful cognitive state is the absolute prerequisite for engaging in the critical reflection necessary for genuine moral and epistemic judgment.

Guided by the framework of Value Sensitive Design[41], the practical viability of this normative solution is in the targeted application of "micro-frictions" that break epistemic dependence without dismantling overall workflow efficiency. This approach is highly applicable to the very domains currently undergoing rapid AI integration, such as computer-assisted translation and digital pedagogical environments. For instance, in contemporary CAT tools (Smartcat, Matecat, or CafeTran etc), the seamless auto-propagation of Neural Machine Translation often "aggressively smooths" over culturally marked lexis and paralinguistic nuances, leading to the semantic homogenization discussed earlier. A robust "friction-by-design" architecture would interdict this seamless flow. Instead of automatically substituting a culturally bound term or a nuanced political idiom with a statistically probable equivalent, the interface would intentionally halt the process, generating a "cognitive checkpoint." The system would make the translator manually select from given options or explicitly confirm the cultural subtext before proceeding. In this way it could reclaim the human role from an "ethical buffer" back to an active cultural author.

[38] Vavouras et al., 2024.
[39] Frischmann & Selinger, 2018.
[40] Kahneman, 2011.
[41] Friedman & Hendry, 2019.

This dynamic is equally visible in digital learning environments like Miro, Lino or Padlet. A completely 'frictionless' AI can instantly evaluate complex analytical tasks or group projects, effectively sidelining the educator and reducing them to a passive spectator of the learning process. Introducing 'cognitive friction' means designing the system to intentionally pause rather than rush to a conclusion. At deeply human junctures – such as interpreting a culturally sensitive perspective or recognizing a neurodivergent student's unique communication style – the software must halt and yield to the teacher's empathetic judgment. The interface would legally mandate the human instructor to explicitly validate or override the algorithm's logic before a final grade is registered. In both professional contexts, these mandatory forcing functions ensure that efficiency is consciously subordinated to the frictional labor of human expertise, proving that the preservation of the autonomous subject is a practically viable design choice.

Furthermore, the transition to prospective accountability, as advocated by Sven Nyholm[42], requires a new legal status for techno-corporate actors. We argue that the "responsibility gap" is not an inevitable feature of machine learning, but a strategic byproduct of corporate opacity and the "black box" nature of proprietary code[43]. Current liability models are often retrospective. They seek to find a single guilty party after harm has occurred. In complex socio-technical systems, this leads to "organized irresponsibility," where the developer, the data provider, and the institutional user each point to the other's role in the automated process.

To bridge this gap, regulatory frameworks must adopt a "relational responsibility model". Due to the relational ethics of Abeba Birhane[44], this model would distribute legal and moral accountability across the entire network of actors involved in the system's lifecycle. This is not simply a "joint liability" clause, but a requirement for "algorithmic traceability". Corporations must be legally required to provide "explainability audits" that translate mathematical weights into human-understandable normative reasons. If an AI system in a university proctoring context flags a student as "suspicious," the institution must be able to present the specific relational data that led to this conclusion. Accountability, in this

[42] Nyholm, 2020.
[43] Pasquale, 2015.
[44] Birhane, 2021.

sense, becomes a precondition for deployment, a "prospective" duty that resides in the design phase, ensuring that the system is "accountable by default" before it ever enters the info sphere.

Finally, international regulation must confront the specter of digital colonialism inherent in the "Brussels Effect." While the EU AI Act is often hailed as a global gold standard, it effectively imposes a Western-centric, individualistic ethical framework on a diverse global population. As analyzed in this work regarding semantic homogenization, the imposition of standardized "ethical guardrails" by a few transnational technology corporations functions as a form of algorithmic governance that bypasses democratic sovereignty.

A truly inclusive regulatory framework for 2026 and beyond must prioritize "epistemic sovereignty". This principle demands that local cultures, linguistic communities, and educational systems have the legal right to define their own boundaries of AI integration. It challenges the "extractivist" model of data collection, where the lived experiences of diverse populations are obtained as "raw material"[45] only to be returned as standardized, hollowed-out algorithmic services. Regulation should therefore protect the "semantic friction" of local languages and the autonomy of local knowledge systems. Without such a robust, philosophically grounded legal intervention, the "responsibility gap" will continue to serve as a shield for corporate impunity, ultimately leading to the total displacement of human subjecthood and the erosion of democratic participation in the age of algorithms.

Discussion

The necessity of fundamentally reconceiving moral accountability is directly conditioned by the global transition toward "AI-mediated societies," a paradigm in which algorithms are no longer simple administrative tools but have become the primary instruments of institutional and political governance. Our analysis suggests that as automated systems increasingly mediate access to education, healthcare, and justice, they generate distinctive forms of vulnerability that traditional ethical frameworks fail to address.

[45] Birhane, 2021.

The most pressing threat identified in this study is the rise of algocracy, where human deliberation is systematically replaced by the data-driven outputs of predictive algorithms.

The empirical evidence presented in this work regarding the "illusion of control"[46] provides a critical neurophysiological foundation for the concept of moral deskilling. This study argues that the decoupling of action and accountability is not merely a psychological byproduct but a structural feature of modern AI interfaces. As Vallor S. and Vierkant T.[47] suggest, the "ethical muscle" of the human agent atrophies when the frictional labor of judgment is delegated to a "black-box" system. We extend this argument by demonstrating that in AI-mediated environments, the subject is effectively transformed into a "moral patient", that is a passive recipient of algorithmic outcomes who lacks the capacity for moral reciprocity. This lack of reciprocity is catastrophic for social trust. As artificial systems are fundamentally invulnerable and cannot experience regret or shame, the relational foundation upon which human institutional trust is built inevitably collapses.

Furthermore, the transition to epistemic dependence marks a radical concentration of power that transcends traditional legal jurisdictions. Our findings regarding semantic homogenization in intercultural communication illustrate how generative AI functions as a "soft power" instrument. By imposing standardized linguistic and ethical frameworks, transnational technology corporations execute a form of digital colonialism that hollowing out local cultural and semantic diversity. This study aligns with Birhane's relational ethics[48], arguing that the prevailing "data-as-oil" metaphor is fundamentally dehumanizing. Data is a profound reflection of human lives and social relations, and its extraction as a "raw resource" strips it of the context necessary for genuine justice. Consequently, the "responsibility gap" identified by Matthias A.[49] must be reinterpreted not just as a technical void, but as an epistemic injustice where the lived realities of marginalized populations are rendered invisible by Western-centric algorithmic patterns.

Addressing this crisis requires a transition away from the "colonial extraction" model toward what we define as prospective

[46] Salatino et al., 2025.
[47] Vallor & Vierkant, 2024.
[48] Birhane, 2021.
[49] Matthias, 2004: 175–183.

accountability. As Nyholm S.[50] emphasizes, society must move beyond retrospective liability and toward an "ethics by design" approach that hardcodes moral safeguards into the sociotechnical infrastructure before deployment. This necessitates a robust political defense of digital sovereignty. We argue that the preservation of human identity in the age of algorithms requires the conscious design of "friction-heavy" interfaces that actively demand human intervention. Without such a philosophical and legislative intervention, we risk a future where human agency is permanently subsumed by proprietary interests, effectively hollowing out the democratic participation that defines the autonomous subject.

Limitations

This analysis is primarily theoretical and philosophical in orientation, relying upon qualitative synthesis rather than direct quantitative measurement of cognitive decline or accountability metrics. The conceptual frameworks discussed regarding functional agency and moral deskilling reflect the scholarly consensus of preliminary within 2025–2026. Taking into consideration the rapid evolution of generative AI capabilities, the phenomenological experience of human-AI interaction may continue to shift, requiring ongoing empirical validation.

Conclusions

This study has demonstrated that the pervasive integration of artificial intelligence into intellectual labor marks a fundamental "relational turn" that challenges the very core of human subjectivity. The ontological analysis conducted herein confirms that while AI agents, such as Large Language Models, exhibit a highly sophisticated simulation of agency, they lack the phenomenal consciousness and authentic intentionality required for genuine moral personhood. The "intentional stance" adopted by users, while pragmatically useful for predicting system behavior, functions as a

[50] Nyholm, 2020.

phenomenological trap that induces a gradual process of moral deskilling.

The empirical and psychophysiological evidence, particularly the neuroimaging data involving EEG[51], reveals a critical "illusion of control" wherein human operators remain neurologically engaged in automated processes while becoming ethically detached from their outcomes. This decoupling, coupled with the phenomena of cognitive offloading and epistemic dependence, accelerates a "verification crisis" in which professional expertise atrophies due to the removal of necessary cognitive friction. In applied domains such as intercultural communication and educational oversight, these processes manifest as semantic homogenization and algorithmic injustice, intensifying the risks of digital colonialism and the concentration of power within techno-corporate infrastructures.

Consequently, traditional retrospective models of responsibility, which focus on assigning individual blame after harm has occurred, are no longer sufficient for governing hybrid sociotechnical systems. This research advocates for a prescriptive shift toward prospective accountability, where ethical safeguards and oversight mechanisms are hardcoded into the system's architecture through a "friction-by-design" approach. The implementation of mandatory cognitive checkpoints and explainability audits is essential to bridge the "responsibility gap" and protect digital sovereignty. Ultimately, preserving human identity in an algocratic society necessitates the conscious design of technologies that supplement rather than replace human interpretive judgment, ensuring that the space for autonomous choice and moral reflection remains the foundation of democratic participation.

References

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), Article 100205. https://doi.org/10.1016/j.patter.2021.100205

Clark, A. (2025). Extending minds with generative AI. *Nature Communications*, 16(1), Article 1–4. https://doi.org/10.1038/s41467–025-59906–9

Czyszczyń, M. (2025). Artificial intelligence as a moral agent: Regulatory implications and a relational–contextual extension of Moor's classification. *Security and Defence Quarterly*, 52(4). https://doi.org/10.35467/sdq/190245

[51] Salatino et al., 2025.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3), 245–268. https://doi.org/10.1007/s13347–015-0211–1

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

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

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Formosa, P., Hipólito, I., & Montefiore, T. (2025). Artificial intelligence (AI) and the relationship between agency, autonomy, and moral patiency. *arXiv*. https://arxiv.org/abs/2504.08853

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press. https://doi.org/10.1017/9781316544846

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), Article 6. https://doi.org/10.3390/soc15010006

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Martela, F. (2025). Artificial intelligence and free will: Generative agents utilizing large language models have functional free will. *AI and Ethics*, 5, 4389–4400. https://doi.org/10.1007/s43681–025-00740–6

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/10.1023/B:ETIN.0000047478.11105.a9

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

Papacharalambous, C. (2021). Other's caress and God's passing by: Levinas encountering Heidegger. *Dia-noesis: A Journal of Philosophy*, 11, 77–94. https://dianoesis-journal.com/wp-content/uploads/2023/09/dianoesis-olot11_opt.pdf

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Salatino, A., Prével, A., Caspar, E., & Lo Bue, S. (2025). Influence of AI behavior on human moral decisions, agency, and responsibility. *Scientific Reports*, 15, Article 12329. https://doi.org/10.1038/s41598–025-95587–6

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.

UNESCO. (2025). *Guidance for generative AI in education and research*. UNESCO Publishing.

Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Philosophy & Technology*, 37(2), Article 54. https://doi.org/10.1007/s13347–024-00742–0

Vavouras, E. (2020). The political philosophy as a precondition and completion of political economy in the *Ways and Means* of Xenophon. *Dia-noesis: A Journal of Philosophy, 9*, 183–198.

Vavouras, E., Koliopoulou, M., & Manolis, K. (2024). From participatory leadership to digital transformation under the interpretation of political philosophy: Types of leadership in education and school administration. *Dia-noesis: A Journal of Philosophy*, 15, 153–170. https://doi.org/10.12681/dia.38171

Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology*, 27(1), Article 14. https://doi.org/10.1007/s10676–024-09800-x

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