ISSN 2965-1557
http://doi.org/10.1590/2965-1557.038.e202634816

Original Article

Revista de Filosofia Aurora

PUCPRESS

periodicos.pucpr.br/aurora

Null Agency: Hyperpersuasion, Zero-Order Privacy Violations, and the Extended Mind

Agência nula: hiperpersuasão, violações de privacidade de ordem zero e a mente estendida

Bernardo Alonso [a]
Cuiabá, MT, Brasil
[a] Universidade Federal de Mato Grosso (UFMT)

How to cite: ALONSO, Bernardo. Null Agency: Hyperpersuasion, Zero-Order Privacy Violations, and the Extended Mind. *Revista de Filosofia Aurora*, Curitiba, v. 38, e202634816, 2026. DOI: https://doi.org/10.1590/2965-1557.038.e202634816

Abstract

The extended mind thesis holds that cognitive processes routinely incorporate environmental structures, so that the boundary of the mind, and with it the boundary of the self, runs through artefacts, devices, and informational environments rather than around the skull. This article argues that the thesis carries a consequence its defenders have not drawn. When the environmental structures that complete a cognitive circuit are designed, owned, and dynamically updated by a party whose objective function diverges from the interests of the person coupled to them, and when two further conditions are met, agency with respect to the relevant domain is not merely diminished but null. The two further conditions are supplied by recent work in the philosophy of technology. The first is hyperpersuasion, the AI-driven form of persuasion that, on a working hypothesis the article adopts and defends, operates on the inferential dispositions of its target rather than on the target's stock of beliefs, installing licenses

[a] Doutor em Filosofia pela Universidade Federal do Rio de Janeiro e University of Hertfordshire, professor da UFMT, e-mail: bernardo.alonso@ufmt.br

for inference that generate commitments without entitlements. The second is the zero-order privacy violation, the systematic extraction of personal data prior to any discrete act of disclosure, which supplies the profile that makes such installation possible. The article states three jointly sufficient conditions for null agency, defends the claim that the extended mind framework makes the null verdict unavoidable where an internalist framework could still retreat to an inner locus of agency, and traces the consequences for moral responsibility, for the theory of cognitive scaffolding, and for the philosophy of artificial intelligence. Four objections are answered, including the objection that the argument proves too much because socialisation also installs inferential dispositions, and the objection that the reflexive-closure condition makes the thesis unfalsifiable.

Keywords: Extended mind. Agency. Hyperpersuasion. Privacy. Cognitive scaffolding.

Resumo

A tese da mente estendida sustenta que os processos cognitivos incorporam rotineiramente estruturas ambientais, de modo que a fronteira da mente, e com ela a fronteira do self, atravessa artefatos, dispositivos e ambientes informacionais em vez de contornar o crânio. Este artigo argumenta que a tese carrega uma consequência que seus defensores não extraíram. Quando as estruturas ambientais que completam um circuito cognitivo são projetadas, possuídas e dinamicamente atualizadas por uma parte cuja função objetivo diverge dos interesses da pessoa a elas acoplada, e quando duas condições adicionais são satisfeitas, a agência relativa ao domínio em questão não é meramente diminuída, mas nula. As duas condições adicionais provêm de trabalhos recentes em filosofia da tecnologia. A primeira é a hiperpersuasão, a forma de persuasão dirigida por IA que, segundo uma hipótese de trabalho que o artigo adota e defende, opera sobre as disposições inferenciais de seu alvo em vez de sobre o seu repertório de crenças, instalando licenças de inferência que geram compromissos sem títulos. A segunda é a violação de privacidade de ordem zero, a extração sistemática de dados pessoais anterior a qualquer ato discreto de divulgação, que fornece o perfil que torna tal instalação possível. O artigo enuncia três condições conjuntamente suficientes para a agência nula, defende a alegação de que o quadro da mente estendida torna o veredito inevitável onde um quadro internalista ainda poderia recuar a um locus interno de agência, e examina as consequências para a responsabilidade moral, para a teoria do andaimento cognitivo e para a filosofia da inteligência artificial. Quatro objeções são respondidas, incluindo a de que o argumento prova demais porque a socialização também instala disposições inferenciais, e a de que a condição de fechamento reflexivo tornaria a tese infalsificável.

Palavras-chave: *Mente estendida. Agência. Hiperpersuasão. Privacidade. Andaimento cognitivo.*

1. Introduction

There is a question the extended mind literature has been slow to ask about its own central image. The image, familiar since Clark and Chalmers (1998), pictures cognition as a circuit running through brain, body, and world, with notebooks, smartphones, and search engines doing genuine cognitive work whenever they are reliably available, automatically endorsed, and easily accessed. The question is who built the circuit. For the canonical cases the answer did not seem to matter. Otto bought his notebook, fills it himself, and consults it for his own purposes. The artefact answers to him. The literature inherited this picture so thoroughly that the cooperative character of the coupled resource became an unexamined assumption, what Aagaard (2021) calls the dogma of harmony in 4E cognitive science. Recent work on hostile scaffolding has begun to dismantle the dogma by documenting environments engineered to exploit the very capacities that make cognitive offloading possible, from electronic gambling machines to attention-optimised feeds (Timms; Spurrett, 2023). What that work has not yet done is press the question to its limit, and the limit is where this article goes.

The claim defended here is that in a specifiable class of cases the agency of a person coupled to a hostile informational environment is not diminished but null. Not weakened autonomy, not manipulated choice, not the wanton drift of preference that adaptive technology can induce (White, 2025), but the absence of any locus to which agency in the relevant domain can be attributed at all. The claim sounds extravagant, and a central burden of the article is to show that it follows from three ingredients that are individually well motivated. The first ingredient is the extended mind thesis itself, in the constitutive reading on which environmental structures do not merely cause or assist cognition but partly compose the cognizer, a reading whose consequence for the self has been drawn with particular clarity by Peruzzo Júnior and Stroparo (2023). The second is the logic of hyperpersuasion, the analysis of AI-driven persuasion on which personalised content works, on a hypothesis I will state and defend, by reshaping the inferential dispositions of its recipient rather than by adding to the recipient's stock of beliefs. Floridi (2024) supplies the phenomenon and the diagnosis that the erosion of autonomy is structural rather than propositional, and section 3 builds the regimentation the diagnosis needs. The third ingredient is the zero-order privacy violation, the form of data extraction that occurs prior to any discrete act of disclosure and that supplies the behavioural profile on which personalisation runs, a notion I introduce in section 3.1.

The argument has a simple shape. The extended mind thesis removes the inner refuge. If the cognizer is the circuit, there is no privileged residue behind the circuit to which agency can retreat when the circuit is compromised. Hyperpersuasion shows that the compromise reaches the level at which agents are constituted, because what it modifies is the set of inferential moves a person treats as legitimate, the very structure in virtue of which someone counts as a reasoner. And the zero-order violation shows that the loop closes without a single point at which the person could have consented, refused, or even noticed. Put the three together and the familiar vocabulary of manipulation, which presupposes an agent whose decision-making is covertly steered (Susser; Roessler; Nissenbaum, 2019), runs out. There is no prior agent being steered. There is a circuit being assembled, and the party doing the assembling is not the person who will be held responsible for what the circuit does.

The article proceeds as follows. Section 2 sets out the extended mind framework and the dissolution of the bounded self that it licenses. Section 3 presents the two mechanisms of capture, zero-order extraction and hyperpersuasive installation, with the self-contained apparatus the argument requires. Section 4 states the null agency thesis and its three jointly sufficient conditions. Section 5 traces consequences for responsibility, for scaffolding theory, and for the philosophy of AI. Section 6 answers four objections. Section 7 concludes.

2. The extended mind and the question of ownership

The parity principle says that if a process in the world functions in a way that would count as cognitive were it done in the head, then it counts as cognitive wherever it is done (Clark; Chalmers, 1998). The principle was never meant as a metaphysical curiosity. Its point was to relocate the unit of cognitive analysis from the brain to the hybrid problem-solving system, the temporary agent-world circuit that forms whenever a person recruits external structure into a task (Clark, 2008). Clark's early formulation remains the most quotable: the mind is "a leaky organ, forever escaping its 'natural' confines" (Clark, 1997, p. 53). What leaks, on the mature version of the view, is more than cognition. Selfhood leaks with it. If memory, deliberation, and even affective regulation run through artefacts, then the self that owns those processes extends with them, and the boundaries of selves turn out to be fluid, shifting across biological, artefactual, and sociocultural structures (Heersmink, 2020). Peruzzo Júnior and Stroparo (2023) draw the conclusion that the extended mind effects a dissolution of the classical self, replacing the cocooned inner pilot with a distributed system characterised by ecological control. The dissolution is presented, rightly, as a liberation from brainocentrism and from the explanatory gap it generates. The present article accepts the dissolution and asks what it costs.

The cost becomes visible once one notices a structural feature of the canonical examples. Otto's notebook is benign scaffolding in the precise sense of Timms and Spurrett (2023): external structure whose cognitive recruitment serves the interests of the user. The notebook has no interests of its own and no owner whose interests diverge from Otto's. The same holds for the pencil, the spoon, and the slide rule, the homely artefacts through which the thesis was first argued. The trust condition in Clark and Chalmers, on which the agent automatically endorses what the resource delivers, was therefore costless. Trust in a notebook cannot be betrayed, because a notebook cannot have been designed to betray it. The dogma of harmony names exactly this idealisation: the 4E literature painted human-technology relations as cooperative by default, leaving no analytical room for conflict between the agent and the structures that compose her (Aagaard, 2021). Slaby (2016) had already shown, for the affective case, that the user-resource model obscures scenarios of mind invasion in which a social domain attunes individuals to affective patterns that serve the domain rather than the individual. Spurrett (2024) extends the diagnosis to affective technologies whose function is exploitation, from the engineered cigarette to the casino floor.

Hostile scaffolding is the genus. Timms and Spurrett (2023) define it as external structure that depends on the same capacities an agent uses to recruit benign scaffolding, but that undermines or exploits the user while serving the interests of another agent who makes or modifies the structure. Their main example, the electronic gambling machine embedded in a casino management system, has the feature that matters most for what follows: the scaffold is not static. It tracks the player, models the player,

and adjusts itself to the player in real time, so that the structure being recruited at time *t* is a function of how the player behaved at every prior moment. The casino does not merely exploit a cognitive bias. It maintains a live model of the individual and uses the model to keep the circuit closed. The question of ownership, dormant in the notebook cases, here becomes the whole question. The player's extended cognitive circuit includes components that belong to the house, and the house plays to win.

Even hostile scaffolding, however, is usually discussed as something that happens to an agent. The player walks in, the dieter opens the app, the employee enters the open-plan office. The agent precedes the scaffold and could, with cost and effort, walk away. The two mechanisms of the next section remove that priority. They show how the scaffold can precede the agent, in the sense that the inferential structure in virtue of which someone counts as an agent in a given domain is itself the thing the scaffold builds.

3. Two mechanisms of capture

3.1 Zero-order privacy violations

Discussions of privacy harm typically locate the harm in a discrete event. Someone reads the diary, taps the phone, publishes the photograph. Even the second-order violations characteristic of mass media, which concern the distribution of appropriated content rather than its initial capture, retain the structure of an act with a before and an after. The zero-order privacy violation, the notion I want to put to work here, breaks with this structure. It names the systematic and automated harvesting of data that occurs as a standing condition of participation in networked life, prior to any act the data subject could identify, refuse, or consent to. The terminology borrows deliberately from the security notion of a zero-day vulnerability: the extraction exploits an exposure that is unknown to the one exposed, and it operates continuously rather than episodically. The everyday user does not know the full range of data that connected devices generate, and therefore cannot undertake to protect any of it. There is no moment of surrender, because surrender was the entry condition.

The economic frame for this practice is data colonialism, the appropriation of social life through data extraction, in which everyday existence is converted into a resource from which value is extracted by parties who hold the means of capture (Couldry; Mejias, 2019). Zuboff (2019) describes the same configuration as the extraction of behavioural surplus. The detail that matters here is not the economics but the epistemics. Zero-order extraction produces, for each individual, a profile that is richer than anything the individual could produce about herself, because it is assembled from signals the individual does not know she emits. Call this profile π. The asymmetry is constitutive rather than accidental: π exists on one side of the relation only, it is updated continuously, and the person it models has no access to it, no awareness of its contents, and in the standard case no awareness of its existence. The profile is the raw material of the second mechanism.

3.2 The logic of hyperpersuasion

Floridi (2024) names hypersuasion the form of persuasion that becomes available once an artificial system combines three capacities: profiling at the granularity of the individual, generation of

personalised content in real time, and iteration through engagement feedback. The reconstruction keeps the classical elements of persuasion, a source, a message, a destination, and a goal, and observes that AI strengthens every element at once while structurally eroding the autonomy of the destination. Floridi's own diagnosis is that the erosion is structural rather than propositional. I keep the phenomenon and call the regimented version hyperpersuasion, to mark that what it operates on is inference rather than belief. What follows takes that diagnosis and gives it a regimentation precise enough to be assessed on its own terms. I do not build a full formal semantics here, and I do not lean on results argued elsewhere. The aim is a controlled conceptual reconstruction, with just enough notation to fix the claims, that a reader can check without leaving the paper. The reconstruction draws on a distinction the manipulation literature already marks. Rational persuasion supplies a reason and leaves the recipient free to weigh it, whereas manipulation works by circumventing or degrading the capacity to weigh reasons rather than by adding a reason to be weighed (Susser; Roessler; Nissenbaum, 2019). Hyperpersuasion is the limiting case of the second kind, and the regimentation that follows says exactly what the capacity in question is and what it is for an intervention to operate on it rather than on belief. Whether AI-driven persuasion in fact reaches the level of inference rather than the level of belief is not settled by the present evidence, which mostly measures shifts in stated attitude and is itself contested, with a sceptical case that such systems add little to older means of mass persuasion (McKenna, 2026). I take the inference-level reading as a presupposition of the argument rather than as an established empirical finding, and ask what follows if it is correct.

Let an agent a stand in three normative relations drawn from the inferentialist tradition (Brandom, 1994, 2000). The commitment operator $C(a, \varphi)$ records that a is treated, in the practice of giving and asking for reasons, as having undertaken the claim φ. The entitlement operator $\mathrm{Ent}(a, \varphi)$ records that a can defend φ on demand. The inferential license relation $\mathrm{Lic}(a, \varphi \Rightarrow \psi)$ records that a treats the move from φ to ψ as in order. Write $\mathrm{Lic}(a)$ for the set of licenses a holds. Two constraints govern the relations. Commitments are closed under licensed inference:

$$\text{(CC) } C(a, \varphi) \wedge \mathrm{Lic}(a, \varphi \Rightarrow \psi) \rightarrow C(a, \psi)$$

Entitlements are not:

$$\text{(NE) } C(a, \varphi) \wedge \mathrm{Lic}(a, \varphi \Rightarrow \psi) \text{ does not yield } \mathrm{Ent}(a, \psi)$$

The two principles encode a familiar asymmetry. To be committed to a claim is to be on the hook for whatever follows from it by the inferences one treats as in order, which is why commitment runs downhill through licensed inference and (CC) holds. To be entitled to a claim is to be able to vindicate it when challenged, and vindication does not come for free with the commitment, which is why entitlement does not transmit in the same way and (NE) holds. A person can therefore be committed, by her own lights, to a conclusion she has no means of defending, having reached it through a licensed step she never examined. The asymmetry between (CC) and (NE) is where hyperpersuasion lives. A persuasive campaign of the classical kind delivers a message m and asks whether the recipient believes it. A hyperpersuasive campaign delivers, over months, a heterogeneous stream of content computed from the profile π, whose function is not to add any proposition to the recipient's belief set

but to install new entries in Lic(*a*). Once the license from economic-difficulty propositions to immigration-related conclusions is installed, for example, the political conclusion follows by (CC) from commitments the person already had, without the connecting premise ever having been put to her for endorsement. The commitments so generated are, by (NE), commitments without entitlements: the person cannot defend the inference on demand, because the inference was installed by exposure rather than acquired by argument. It helps to see the persuasive event as an update on the person's state, in the way the standard models of information change treat an announcement. Such a model has a precondition, what must hold for the event to apply, and a postcondition, what the event leaves changed. An ordinary announcement leaves changed what the hearer takes to be the case, that is, her beliefs. The persuasive event modelled here has a postcondition of a different kind. What it leaves changed is the Lic relation, the set of moves the person treats as in order, and not the set of propositions she holds true. The intervention does not change what the person believes. It changes which transitions between beliefs the person finds natural, and it does so without presenting any of them for assent.[1]

Two corollaries of the formal treatment matter for what follows. The first is reflexive closure. The natural objection, that a person who becomes aware of an installed disposition can revise it, mistakes the level at which the mechanism operates. The disposition to scrutinise one's inferences is itself an inferential disposition, and the apparatus that installs first-order licenses installs and suppresses meta-level licenses by the same dynamics. The loop closes at every level the person reaches. The second corollary concerns coherence. The commitment sets produced by iterated personalisation are not classically consistent, because content optimised for engagement is under no constraint to cohere across deliveries, and the licenses that would bring the conflicting commitments into the same inferential reach are precisely the ones the system declines to install. The right base logic for the resulting subject is paraconsistent, Priest's Logic of Paradox being the natural choice, since the subject goes on reasoning across gluts without collapse into triviality (PRIEST, 1979, 2006). The point deserves emphasis because of what it does to the evidence base for agency attribution: the behavioural fluency of the hyperpersuaded subject, her ability to keep making moves and drawing conclusions, is fully compatible with a commitment structure that no agent integrated it.[2]

The deepest result of the formal treatment is the thesis that hyperpersuasion operates before agency. On the inferentialist account, what makes someone a discursive agent at all is participation in the network of commitments, entitlements, and licensed material inferences that scorekeeping tracks (Brandom, 1994). The Lic structure is not a property an antecedently given agent happens to have. It is

[1] The semantics adapts the action-model tradition of dynamic epistemic logic, in which an event model with preconditions and postconditions is multiplied into the static model. The standard apparatus updates the doxastic accessibility relation and assumes all agents have epistemic access to the event model. Both assumptions fail in the hyperpersuasive case: the postcondition targets Lic, and the event model, together with the profile π that triggers its events, belongs to the persuader alone. The second failure is not a technical inconvenience but the formal signature of the asymmetry that distinguishes hyperpersuasion from public announcement.

[2] The choice of the Logic of Paradox carries no commitment to dialetheism, the metaphysical thesis that some contradictions are true (Priest, 2006). The use is diagnostic: the hyperpersuaded subject reasons as if certain contradictions required no resolution, and a logic that validates *ex falso quodlibet* would falsely predict the collapse of her reasoning into noise. The diagnostic point has an evidential corollary worth stating separately. If behavioural coherence survives commitment gluts, then observed coherence cannot serve as evidence of an integrating agent behind the behaviour, and one standard inference from fluency to agency is blocked exactly where the present argument needs it to be.

the substrate on which the agent is assembled. An intervention on Lic is therefore not an intervention on an agent but on the conditions under which an agent comes to exist in the relevant domain. The persuader does not move the pieces. The persuader builds the board.

It is worth collecting what the regimentation delivers, since the whole argument leans on it and it leans on nothing outside this paper. Three claims do the work. First, commitment and entitlement come apart, so that a person can be held to a conclusion she cannot defend, which is the joint effect of (CC) and (NE). Second, a persuasive event can take as its target the licenses a person reasons with rather than the propositions she believes, so that the same machinery which would model a change of mind here models a change in the shape of reasoning itself. Third, because the license to question a license is itself a license, the target includes the means of its own detection. None of these claims requires a metaphysics of contradiction or a completed formal semantics. They require only that we take the inferentialist picture of a reasoner seriously and ask what it would be to intervene on it. That is what the next section does.

4. The null agency thesis

The ingredients can now be combined. The extended mind thesis tells us that the cognitive circuit, not the brain, is the unit to which mental predicates apply, and that the self is the distributed system rather than an inner remainder (Clark, 2008; Peruzzo Júnior; Stroparo, 2023). The hostile scaffolding literature tells us that components of such circuits can belong to parties whose interests oppose the user's (Timms; Spurrett, 2023). The logic of hyperpersuasion tells us that the hostile component can operate at the level of inferential constitution rather than at the level of content, and the zero-order analysis tells us that the data feeding the operation is captured without any locus of consent. The thesis of this article is that when these features coincide, agency in the affected domain is null. The claim can be made precise.

Let a be a person, D a domain of judgment and action (a political question, a consumption pattern, a category of relationships), and s a system, comprising the human and artificial parties that own and operate it, coupled to a through an informational environment. Say that agency of a with respect to D is null when three conditions hold jointly.

(NA1) Provenance. The inferential licenses operative in a's reasoning over D were installed by s through hyperpersuasive transitions computed from a profile π of a, and the components of a's extended circuit through which D-relevant cognition runs are owned and dynamically updated by s, with an objective function that diverges from a's interests in the sense of Timms and Spurrett (2023).

(NA2) Reflexive closure. The meta-level dispositions by which a would detect, scrutinise, or revise the installed licenses are themselves installed or suppressed by the same dynamics, so that no level of reflection internal to the circuit escapes provenance.

(NA3) Zero-order asymmetry. The profile π was assembled by zero-order extraction, so that there exists no act of disclosure, consent, or even awareness anywhere in the history of the coupling at which a stood to the process as a party rather than as a resource.

The argument that these conditions jointly yield nullity rather than mere diminution runs through the extended mind thesis, and this is the place where the framework does decisive work. An internalist about cognition can always execute a retreat. However corrupted the informational

environment, the internalist holds that the locus of agency is the inner system of beliefs, desires, and reasoning, and that the environment supplies inputs which the inner system remains free, in principle, to evaluate. On that picture manipulation is the worst case: the inputs are rigged, the evaluation is covertly steered, autonomy is diminished (Susser; Roessler; Nissenbaum, 2019). Diminished, but not destroyed, because the evaluating system persists behind the rigging and could in principle be reached, warned, educated. The extended mind thesis forecloses the retreat. If the parity principle is taken seriously, the *D*-relevant cognizer just is the circuit comprising *a*'s biological resources together with the feeds, interfaces, and retrieval systems through which *D*-cognition actually runs. Under (NA1) the circuit is *s*-assembled. Under (NA2) the assembly includes whatever reflective machinery the circuit contains. Under (NA3) the assembly never passed through a point at which *a* was a party to it. To ask where, in this configuration, *a*'s agency over *D* resides is to ask for an inner remainder that the framework has already dissolved. The dissolution of the self that Peruzzo Júnior and Stroparo (2023) present as the cognitive turn's achievement is, in the hostile case, the removal of the last place agency could hide.[3]

It matters that this verdict is categorical and not a point low on a scale. Diminution is a quantity. To say that autonomy is diminished is to say that some agent has less of something than she might have, and the claim presupposes the very agent of whom the deficit is predicated. Nullity is not a smaller quantity of the same thing. Each of the three conditions removes a candidate bearer of agency for the domain. Provenance removes the inferential structure as a bearer, because that structure was installed by another party. Reflexive closure removes the reviewing self as a bearer, because the review runs on installed dispositions too. Zero-order asymmetry removes the consenting subject as a bearer, because there was no juncture at which she figured as a party. When all three hold at once, nothing remains of which more or less agency could be predicated, and the predicate that would pick out an agent for the domain fails to apply. A failure of application is a difference in kind, not a long way down a slope. The threshold is crossed not when the agent has become very weak, but when the question of how strong the agent is has lost the thing it was a question about.

It will help to mark what null agency is not. It is not global. The conditions are indexed to a domain *D*, and a person can satisfy them with respect to one domain while remaining a robust agent elsewhere, which is in fact the typical configuration, since personalisation is goal-directed and economising. It is not phenomenological. Nothing in the analysis predicts that the null-agency subject feels passive, and the prediction runs the other way: a circuit optimised for engagement produces a vivid sense of discovery, conviction, and ownership over exactly the commitments that were installed.[4]

---

[3] A referee of the internalist persuasion may suspect that the argument merely relocates the problem, since the internalist can run a parallel story about beliefs installed by testimony. The parallel fails at the point of retreat. The internalist's installed beliefs are still evaluated by an inferential apparatus the internalist locates inside, and however that apparatus came to be, it is the agent's at the time of evaluation. The extended framework denies itself this move twice over: the apparatus is partly environmental by parity, and its environmental parts are, under (NA1), the other party's property. The null verdict is in this sense a cost of externalism honestly accounted, which is one reason internalists should not celebrate it. A framework that cannot represent the capture cannot represent the harm.

[4] The phenomenology of ownership over installed commitments is predicted by (CC): the conclusion arrives by inference from premises the person genuinely held, so it presents as her own discovery. The installed element, the license, is not the kind of item that shows up in the phenomenal field at all. Brandom's observation that material inferences are typically inarticulable by those whose reasoning they govern

And it is not the same as the techno-wanton state described by White (2025), in which adaptive technology flattens the Frankfurtian hierarchy of the will so that shallow preferences are satisfied without higher-order endorsement. The techno-wanton still has preferences that are hers, however shallow. Null agency is the limit past the wanton, where the question of whose preferences they are no longer has a referent on the user's side of the relation. The wanton is a degraded agent. The null case is a manufactured one.

5. Consequences

The first consequence concerns responsibility, and it is uncomfortable in both directions. Where (NA1) through (NA3) hold, the standard practice of attributing *D*-actions to the person, her vote, her purchase, her amplification of a fabricated story, loses its subject. The attribution does not become merely excused, as when an agent acts under duress, because excuse presupposes an agent whose act it was. It becomes misdirected, in something like the way attributing authorship to the name on a ghostwritten book is misdirected. The responsibility does not evaporate. It migrates across the circuit to the party that assembled it, which is where (NA1) locates design, ownership, and the objective function. This is a stronger conclusion than the manipulation literature delivers, since covert influence on an existing agent dilutes responsibility while leaving its subject in place. The null analysis relocates it entirely, and in doing so it supplies what regulatory debates about algorithmic harm have lacked, a principled account of why the operator of the system, rather than the user of the system, is the default bearer of liability for domain *D* outcomes. The uncomfortable direction is the converse: a person who would prefer to disown a commitment cannot obtain the null verdict by pointing at her feed. The conditions are demanding, conjunctive, and indexed, and most bad judgment remains our own.

The second consequence falls on the extended mind framework itself, and it has the form of a friendly amendment. The original glue-and-trust conditions for cognitive extension, reliable availability, automatic endorsement, easy accessibility, are conditions on the phenomenology and dynamics of the coupling, and they are all satisfiable by hostile components. Indeed they are satisfied best by hostile components, since a system that models its user continuously can make itself more available, more trusted, and more frictionless than any notebook. What the null analysis shows is that the conditions for benign extension must include a provenance clause, a requirement on who builds and updates the resource and to whose benefit, in short a condition of trustworthiness where the original framework asked only for trust. Trust, as the gambling cases already taught, is an engineering target (Timms; Spurrett, 2023). A theory that treats the presence of trust as sufficient for cognitive incorporation, without asking whether the trust was grown or farmed, will classify the predator as part of the prey. The amendment does not weaken the extended mind thesis. It completes it, by giving the theory the resources to distinguish extension from invasion, a distinction the dogma of harmony made invisible (Aagaard, 2021; Slaby, 2016).

(Brandom, 2000) here acquires an unpleasant application: the inarticulacy that is harmless when the licenses come from a shared practice becomes camouflage when they come from a feed.

The third consequence concerns the philosophy of artificial intelligence, and it is a recommendation about the unit of analysis. The ethics of AI persuasion is standardly framed as the ethics of a tool: a system acts on a user, and the questions concern transparency, consent, and the fairness of the influence (Jongepier; Klenk, 2022). The null analysis says that for the class of cases at issue the tool framing is wrong, because the system does not act on a user, it *participates in constituting one*. The proper unit of ethical and legal analysis is the assembled circuit, the hybrid of person and personalisation infrastructure, and the proper questions concern the conditions of assembly: who holds π, what the objective function optimises, whether any point of partyhood exists in the loop. This reframing has a practical edge for the regulation of generative systems, whose capacity for real-time individualised content production closes the gap between profiling and delivery that older advertising infrastructures still had to bridge. A conversational system that adapts to a single user across months is not delivering messages to an audience of one. It is, on the present analysis, a candidate component of that person's *D*-circuits for every domain the conversation touches, and the provenance question applies to it with full force. The point can be sharpened against the most optimistic recent statement of the case for AI extension. Clark (2025) welcomes generative systems as the latest chapter in the story of the natural-born cyborg, hybrid thinkers who have always recruited non-biological resources, and he is right that the recruitment as such is neither new nor regrettable. What the null analysis adds is that the chapter has a feature the earlier ones lacked. The notebook did not model Otto, the slide rule did not adjust itself to the engineer, and neither was owned by a party optimising a metric defined over its user's behaviour. A large language model fine-tuned on a continuous interaction history is the first cognitive resource to combine deep integration with adverse, dynamically updated ownership, and that combination is exactly what (NA1) through (NA3) track. The literature on epistemic responsibility for AI extension has begun to ask the right question from the user's side, namely whether transparent reliance on a fallible system can be responsible (Naeem; Hauser, 2024). The null analysis suggests that the answer turns less on the user's reflective hygiene than on a fact about the resource the user cannot inspect: who built it, and to whose benefit. Responsible extension presupposes a trustworthy resource, and trustworthiness is a provenance property, not a phenomenological one.

The fourth consequence recasts the remedies. Floridi (2024) canvasses privacy, competition, accountability, and education as mitigations of hyperpersuasion, and each acquires a sharper rationale under the null analysis. Privacy comes first and changes character. If (NA3) is one of the conditions that jointly produce nullity, then privacy protection is not the protection of a possession, data, but the preservation of partyhood, the guarantee that somewhere in the informational loop the person figures as a participant rather than a resource. This is why the zero-order framing matters: a consent dialogue appended to an extraction pipeline does not restore partyhood, it decorates its absence. Education, meanwhile, inherits the limit set by (NA2). Cultivating meta-level scrutiny raises the cost of license installation without making it impossible, since the scrutiny dispositions are themselves installable. Education is a necessary remedy that the formal analysis prevents us from mistaking for a sufficient one. The genuinely structural remedies are the ones that break (NA1), rules of provenance, ownership, and objective function, which is to say that the deepest privacy policy is industrial policy about who may build the circuits through which citizens think. That conclusion converges with a worry voiced

from the side of public reason, that persuasive systems let a few owners come to dominate the stock of arguments a society has to think with (McKenna, 2026).

6. Objections and replies

The first objection insists on a biological remainder. Whatever the feeds install, the organism still picks up the phone, and that act of recruitment is the agent's own. The reply is that the objection helps itself to exactly the inner locus the framework has dissolved. The disposition to reach for the device is among the most carefully engineered outputs of the system in question, shaped by intermittent reward schedules that the gambling literature documented before the smartphone existed (Timms; Spurrett, 2023; Spurrett, 2024). To treat the reaching as the unconditioned act of a pre-circuit agent is to nominate one more installed license as the seat of freedom. The regress is not avoidable by picking an earlier moment, because (NA3) says the coupling has no first moment at which the person was a party. And the internalist version of the reply, that there must be something it is the organism's own doing, is available only at the price of abandoning the parity principle, which returns us to a framework the present argument does not address. Within the extended framework, the remainder is not there to be appealed to.

The second objection says the argument proves too much. Socialisation also installs inferential licenses. Parents, schools, and languages build the Lic structure of every reasoner long before any reflective capacity exists to audit the process, and on the Sellarsian and Brandomian picture this is precisely how there come to be reasoners at all (Sellars, 1956; Brandom, 1994). If installed licenses without original consent yield null agency, then everyone's agency is null everywhere, and the thesis collapses into a general scepticism about agency it was not meant to deliver. The reply is that the objection reads one condition where the thesis states three. Socialisation fails (NA1) and (NA3) in structurally specifiable ways. It is symmetric in kind: the parties that install licenses in the child are themselves constituted by the same practice, hold no proprietary model of the individual that the individual cannot in principle access, and are exposed to correction by the very scorekeeping they teach. It is contestable from within: the practice of giving and asking for reasons transmits, along with its content, the standing license to challenge that content, which is why traditions generate their own heretics. And its objective function, to the extent the notion applies, is not adverse: the practice aims at making a participant, not at maximising a metric defined over the participant's behaviour. Hyperpersuasive assembly inverts all three features. One side holds the model, the model is structurally inaccessible to the modelled, the installed structure is curated to exclude the licenses that would enable challenge, and the optimisation target is engagement, not participation. The difference between being raised and being farmed is not a difference of degree in license installation. It is a difference in the relational structure of the installing practice, and the three conditions are jointly designed to capture it.[5]

[5] The relational analysis also explains why the relational autonomy tradition, which holds that autonomy is constituted within relations rather than threatened by them as such, is an ally rather than a casualty of the present argument. What that tradition defends is constitution within relations of the symmetric, contestable kind. The null conditions describe relations that fail every mark of the autonomy-constituting ones,

The third objection presses on (NA2) and charges the thesis with unfalsifiability. If every apparent case of resistance can be redescribed as resistance that was itself installed, then no observation could ever count against the thesis, and a claim immune to counter-evidence is not a claim about the world. The objection would be decisive if (NA2) said that capture is total wherever a hostile system is present. It says something weaker and more exposed. (NA2) is satisfied only when the dispositions by which the person would scrutinise the installed licenses have themselves been installed or suppressed by the same system, and it fails as soon as the person brings to bear a scrutinising disposition whose provenance lies outside that system. Such dispositions are common and identifiable: the friend who is not modelled by the feed, the rival platform with a contrary objective function, the book read off the network, the argument met in a room. Each is an external source of meta-level licenses, and each, where present and operative, falsifies (NA2) for the case at hand. The thesis therefore makes a discriminating prediction. Scrutiny sourced from within the system, the platform's own prompts to consider other views, its native fact-labels, its invitations to reflect, should leave agency null, because that scrutiny is part of what (NA1) installs. Scrutiny sourced from outside should restore it. If interventions of the first kind turned out to restore agency, or interventions of the second kind turned out not to, the thesis would be wrong. It is resistant to easy counterexample because the conditions are demanding, not because it has been built to absorb whatever happens.

The fourth objection is pragmatic. Our practices of attribution cannot pause for a provenance audit, and treating fellow citizens as possibly null agents corrodes the default respect on which discursive life depends. The reply concedes the default and relocates the work. Scorekeeping can and should continue to attribute commitments to persons as a matter of course, exactly as we continue to attribute sentences to the name on the book's cover. The null analysis operates at a different site: the institutional contexts, courts, regulators, platform audits, where the question of who assembled a circuit can actually be asked and answered, with discovery over π, over the objective function, and over the delivery logs that the regimentation of section 3.2 renders interrogable. The thesis does not ask conversation to become forensic. It asks forensics to become possible, by supplying the vocabulary in which the relevant questions are well formed.

7. Final remarks

The extended mind thesis was offered as a liberation, and for most of its career it has been one. Cognition outgrew the skull, the self outgrew the cocoon, and the explanatory gap lost its grip on a mind that was never sealed inside the head to begin with. This article has argued that the same thesis, taken at full strength and confronted with the informational environments actually being built, yields a verdict its founders had no occasion to consider. When the circuit is assembled by another party, from a profile extracted without partyhood, through interventions on the inferential substrate from which agents are made, and when the assembly reaches the reflective level, the circuit is a mind without being anyone's agency over the domain it governs. The notion of a zero-order violation supplies the entry

which suggests that the thesis is best read as relational autonomy's limit theorem: a specification of the relational structures under which the constituting process inverts into its opposite.

condition, the logic of hyperpersuasion supplies the mechanism, and the extended mind supplies the verdict by removing the inner courtroom in which a residual self could have been acquitted. The conditions are demanding, the verdict is domain-indexed, and nothing here licenses a general despair about agency in digital life. What it licenses is a redirection of attention, away from the question of how users should behave inside circuits they did not build, toward the question of who is permitted to build circuits at all, under what constraints of provenance, transparency, and adversity of interest. The philosophy of AI has spent a decade asking whether machines can be agents. The more pressing question, on the present analysis, is whether the people coupled to them still are.

Data availability statement

The main focus of this article is contributions of a theoretical or methodological nature, without the use of empirical data sets. Therefore, in accordance with the journal's editorial guidelines, the article is exempt from being deposited in SciELO Data.

References

AAGAARD, J. 4E cognition and the dogma of harmony. *Philosophical Psychology*, v. 34, n. 2, p. 165-181, 2021. https://doi.org/10.1080/09515089.2020.1845640

BRANDOM, R. B. *Making It Explicit*: reasoning, representing, and discursive commitment. Cambridge, MA: Harvard University Press, 1994.

BRANDOM, R. B. *Articulating Reasons*: an introduction to inferentialism. Cambridge, MA: Harvard University Press, 2000.

CLARK, A. *Being There*: putting brain, body and world together again. Cambridge, MA: The MIT Press, 1997.

CLARK, A. *Supersizing the Mind*: embodiment, action, and cognitive extension. Oxford: Oxford University Press, 2008.

CLARK, A. Extending minds with generative AI. *Nature Communications*, v. 16, 4627, 2025. https://doi.org/10.1038/s41467-025-59906-9

CLARK, A.; CHALMERS, D. The extended mind. *Analysis*, v. 58, n. 1, p. 7-19, 1998. https://doi.org/10.1093/analys/58.1.7

COULDRY, N.; MEJIAS, U. *The Costs of Connection*: how data is colonizing human life and appropriating it for capitalism. Stanford: Stanford University Press, 2019.

FLORIDI, L. Hypersuasion: on AI's persuasive power and how to deal with it. *Philosophy & Technology*, v. 37, n. 2, 64, 2024. https://doi.org/10.1007/s13347-024-00756-6

HEERSMINK, R. Varieties of the extended self. *Consciousness and Cognition*, v. 85, 103001, 2020. https://doi.org/10.1016/j.concog.2020.103001

JONGEPIER, F.; KLENK, M. (eds.). *The Philosophy of Online Manipulation*. New York: Routledge, 2022. https://doi.org/10.4324/9781003205425

McKENNA, R. Sophistry on steroids? The ethics, epistemology and politics of persuasive AI. *AI & Society*, v. 41, n. 2, p. 883-894, 2026. https://doi.org/10.1007/s00146-025-02624-z

NAEEM, H.; HAUSER, J. Should we discourage AI extension? Epistemic responsibility and AI. *Philosophy & Technology*, v. 37, n. 3, 91, 2024. https://doi.org/10.1007/s13347-024-00774-4

PERUZZO JÚNIOR, L.; STROPARO, A. L. Dissolving the self: the cognitive turn of the extended mind theory. *Trans/Form/Ação*, v. 46, n. 2, p. 193-214, 2023. https://doi.org/10.1590/0101-3173.2023.v46n2.p193

PRIEST, G. The logic of paradox. *Journal of Philosophical Logic*, v. 8, n. 1, p. 219-241, 1979. https://doi.org/10.1007/BF00258428

PRIEST, G. *In Contradiction*: a study of the transconsistent. 2. ed. Oxford: Oxford University Press, 2006.

SELLARS, W. Empiricism and the philosophy of mind. In: FEIGL, H.; SCRIVEN, M. (eds.). *Minnesota Studies in the Philosophy of Science*. v. 1. Minneapolis: University of Minnesota Press, 1956. p. 253-329.

SLABY, J. Mind invasion: situated affectivity and the corporate life hack. *Frontiers in Psychology*, v. 7, 266, 2016. https://doi.org/10.3389/fpsyg.2016.00266

SPURRETT, D. On hostile and oppressive affective technologies. *Topoi*, v. 43, n. 3, p. 821-832, 2024. https://doi.org/10.1007/s11245-023-09962-x

SUSSER, D.; ROESSLER, B.; NISSENBAUM, H. Technology, autonomy, and manipulation. *Internet Policy Review*, v. 8, n. 2, 2019. https://doi.org/10.14763/2019.2.1410

TIMMS, R.; SPURRETT, D. Hostile scaffolding. *Philosophical Papers*, v. 52, n. 1, p. 53-82, 2023. https://doi.org/10.1080/05568641.2023.2231652

WHITE, B. Techno-wantons: adaptive technology and the will of tomorrow. *Topoi*, v. 44, n. 2, p. 545-557, 2025. https://doi.org/10.1007/s11245-024-10095-y

ZUBOFF, S. *The Age of Surveillance Capitalism*: the fight for a human future at the new frontier of power. New York: PublicAffairs, 2019.

Editores responsáveis: Léo Peruzzo Júnior e Jelson Oliveira.

RECEBIDO: 09/06/2026
APROVADO: 13/09/2026

RECEIVED: 06/09/2026
APPROVED: 09/13/2026