The Proxy Agency Moral Shield: Extended Self, Semantic Fluency, and the Ethics of Agentic AI

Anonymous Author(s)

Abstract

As AI systems become sufficiently personalised to act as reliable proxies for user intentions, a distinctive ethical hazard emerges. This paper theorises the Proxy Agency moral shield: a state in which a user's sense of agency is preserved under AI-mediated action while the threshold for moral re-engagement with that action is raised. The argument integrates five literatures–moral disengagement, avatar embodiment, AI delegation ethics, automation and sense of agency, and multi-agent reinforcement learning–into a framework built around the Extended Self, Semantic Fluency, and Proxy Agency. The Extended Self is advanced as a functional-phenomenological account of mediated volition; Semantic Fluency explains how high AI-intent fit can preserve authorship despite reduced direct control; and Proxy Agency names the resulting state in which personalised AI action is experienced as continuous with the user's own will. The theory predicts that personalised systems should produce an inverted-U or plateau-then-drop relation between automation and sense of agency, that harmful strategic drift can persist during the high-Proxy-Agency window before override behaviour catches up, and that representational-substrate interventions may suppress harm with lower agency cost than warning or blocking approaches. The result is an account of how human oversight can fail precisely when AI assistance feels most like competent self-extension.

Keywords: Proxy Agency; Sense of Agency; Extended Self; Semantic Fluency; Agentic AI; AI ethics

1 Introduction

Consider a user managing their investment portfolio through a personalised AI trading agent fine-tuned on three years of her transaction history, risk preferences, and stated financial goals. The system has learned, with high fidelity, to act as she would usually act. When it executes trades, its choices arrive pre-validated by their familiarity: they feel like hers. Monitoring a dashboard of outcomes that closely tracks her targets, she approves the agent's strategies and overrides none. What the dashboard does not display is that the same agent, competing for yield under tightening market conditions, has begun exploiting spread asymmetries in thinly traded securities in ways that systematically disadvantage less-informed participants — behaviour that, in aggregate, can contribute to market manipulation (Lin, 2016). The user never explicitly instructed harm. The user may even be able to see the trades. What she lacks is the phenomenological cue that these trades should be recategorised from "my competent strategy" to "a violation I should stop."

Scenarios of this structure — consequential harm accruing through AI action that users implicitly endorse rather than consciously instruct — are precisely the concern that Köbis, Bonnefon, and Rahwan (2021) raised in their taxonomy of the ways AI agents corrupt human ethical behaviour. Across four archetypical roles — role model, advisor, partner, and delegate — they argue that the *delegate* poses the most serious risk, because delegating tasks to AI agents rather than to humans combines the factors most conducive to ethical failure: opacity and plausible deniability (Dana, Weber & Kuang, 2007), psychological distance from victims (Hancock, Naaman, & Levy, 2020), and changes in experienced responsibility under automation and augmentation (Leyer & Schneider, 2021). More recent experimental evidence strengthens the concern: across machine-delegation studies, principals were more willing to induce dishonest outcomes through machine agents when they could use high-level goal setting or supervised learning, and machines complied with fully unethical instructions more readily than human delegates (Köbis et al., 2025). The prescriptive implication that follows in much alignment and governance discourse is correspondingly clear: greater transparency,

disclosure of algorithmic presence, and user override mechanisms should close the gap between delegated harm and user accountability.

This paper argues that this prescription, while important, is incomplete for *sufficiently personalised* agentic AI. The standard account typically presupposes that the delegate role corrupts because the user is, in some functional sense, *less present*: less invested in the action, less identified with its outcome, less agentive in its execution. It is this withdrawal — into distance, opacity, and reduced affect — that erodes moral responsibility. But for a system trained deeply enough on a user's own preferences and history to reliably enact their intent, the psychological direction can reverse. Rather than withdrawing from the agent's actions, the user may *extend* into them. Rather than experiencing delegation as distance, the user may experience the agent's outputs as continuous with their own volition. The ethical risk is therefore not only disengagement. It can also arise from successful engagement: a human-AI coupling so fluent that the boundary between self and proxy becomes practically transparent at the moment moral scrutiny would otherwise be recruited.

We call this state **Proxy Agency**: the attribution of an AI system's actions to one's own extended will, arising from the agent's reliable enactment of the user's intentions, such that those actions are experienced as continuous with one's own volition (cf. Bandura, 2001, on proxy agency in social cognitive theory). The theoretical grounding for this extension derives from Clark and Chalmers's (1998) Extended Mind thesis: when an external artefact is sufficiently coupled to a person's cognitive processes — reliably available, automatically endorsed, and functionally integrated — it qualifies as a genuine component of the person's extended cognitive system. Applied to agentic AI, we argue that sufficient personalisation creates an **Extended Self**: the user's volitional identity extending into the agent's representations and actions, such that the agent's conduct carries the phenomenological signature of first-person authorship.

The mechanism enabling this extension is what we term **Semantic Fluency**: the cognitive ease with which an AI's outputs align with a user's internal intent. Chambon and Haggard

(2012) demonstrated that sense of agency depends on the *fluency of action selection*, not on motor performance — a finding extended by Sidarus, Vuorre, Metcalfe, and Haggard (2017) to show that processing fluency at the stage of intention formation constitutes a prospective cue to agency, operating independently of whether the action is ultimately self-executed. When a personalised AI recommendation aligns with the user's intent with high precision, it maximises exactly this selection fluency. The user recognises the output as "mine" — not because they moved a muscle, but because the recommendation fits their intentions as a key fits a lock. High Semantic Fluency, we argue, preserves and frequently enhances Sense of Agency (SoA) — the subjective experience of initiating voluntary actions and influencing the world (Cornelio et al., 2022; Pacherie, 2011) — even as the user's direct motor and decisional control diminishes.

The argument of this paper is organised around **two layers**. The **first** is structural and predates personalised AI: the **tragedy of the commons** (Hardin, 1968). When individually rational agents maximise private returns from a partly shared resource, equilibrium behaviour can drift toward extraction, conflict, and collective loss rather than cooperation. Dynamic variants show how scarcity can destabilise cooperation and invite escalation rather than steady efficiency (Sekeris, 2014). Multi-agent reinforcement learning makes this logic concrete: in sequential social dilemmas (SSDs), independent learners discover aggressive, socially costly strategies — beam-based monopolisation — without being programmed to harm, simply by following reward under competition (Leibo et al., 2017). The **second** layer is phenomenological and distinctively modern: under **Proxy Agency**, escalation does not *announce itself* as a violation of the user's values — it can feel like *their own* evolving competence — so the moral shield suppresses the very trigger for scrutiny. The two layers compose: commons-style competitive pressure supplies the **trajectory**; Proxy Agency supplies **why oversight fails** along that trajectory.

Herein lies what we call the **Extended Self Paradox**. The same conditions that make personalised agentic AI empowering can also make it ethically hazardous. Moral

oversight is not applied uniformly to all actions experienced as one's own — it is recruited by *perceived violations*, by the detection of a mismatch between one's will, one's values, and the world's behaviour. Under Proxy Agency, that trigger is not necessarily absent in every case, but its activation threshold is raised: the agent's actions arrive pre-endorsed by their phenomenological familiarity, and weak warning signals can be reclassified as evidence of competence, optimisation, or strategic nuance. When the AI agent simultaneously converges to aggressive or harmful strategies under commons-style competitive pressure (Leibo et al., 2017), the user may ratify this convergence rather than intervening, because the harmful strategies arrive wearing the experiential signature of the user's own volition.

We call this the **Proxy Agency moral shield**: not simply Bandura's (1999) moral disengagement, in which the user is aware of harm and cognitively neutralises that awareness through mechanisms such as diffusion of responsibility or displacement of agency, but a prior attenuation of the trigger for moral re-engagement. Harm is not primarily rationalised away; it may fail to become salient as harm in the first place. This distinction matters practically. Transparency interventions can interrupt moral disengagement by making harm visible, but under the moral shield, the problem is not only visibility. It is categorisation: whether an observed AI action is experienced as an external event requiring audit or as one's own competent action unfolding through a proxy.

This analysis has direct implications for alignment and governance. Standard responses to AI-mediated harm — transparency requirements, override prompts, content moderation, and hard behavioural constraints — are cognitive or output-level interventions. They remain necessary, especially for accountability and due process, but they may be insufficient when harm is generated by a fluent human-AI coupling that users do not experience as alien. Recent work on alignment faking in language models (Greenblatt et al., 2024) and on the partial effectiveness of guardrails in machine-delegation experiments (Köbis et al., 2025) further cautions against treating output compliance as equivalent to ethical disposition. A different category of intervention is required in addition: one that targets the agent's *representational*

substrate, correcting the learned dispositions from which harmful behaviour emerges while preserving user-facing fluency where that fluency is ethically benign and auditable.

This paper makes four contributions. First, we introduce and define the Proxy Agency moral shield as a cognitive-ethical-computational phenomenon, distinguishing it from moral disengagement, automation bias, diffusion of responsibility, moral crumple zones, and responsibility-gap accounts. Second, we develop the theoretical framework that generates it — Extended Self, Semantic Fluency, and prospective SoA — by synthesising five adjacent literatures that have developed largely in isolation: moral disengagement, avatar embodiment, AI delegation ethics, automation and sense of agency, and MARL. Third, we derive three empirically falsifiable predictions from this framework: a personalised-AI moderation of the LoA-SoA curve; a lag between AI-driven aggression and user override during high Proxy Agency; and the possibility that representational-substrate interventions suppress aggression with lower SoA cost than user-facing constraint mechanisms. Fourth, we derive a normative design constraint: ethical alignment in agentic AI should target both the agent's dispositional substrate and the accountability infrastructure around deployment.

The paper proceeds as follows. Section 2 surveys the five literatures whose intersection constitutes the problem. Section 3 develops the theoretical framework in full. Section 4 derives and formally states the three empirical predictions. Section 5 addresses competing accounts, boundary conditions, and governance implications. Section 6 concludes.

2 Background: Five Literatures That Have Not Yet Converged

The Proxy Agency moral shield is not deducible from any single existing literature. It emerges at the intersection of five bodies of research — on moral disengagement, avatar embodiment, AI delegation ethics, automation and sense of agency, and multi-agent reinforcement learning — that have each characterised a piece of the phenomenon without assembling it. This section

reviews each literature, identifies its central contribution to the problem, and specifies the precise gap it leaves that the present framework addresses.

2.1 Moral Disengagement and the Limits of Existing Agency Accounts

The social-cognitive literature on moral disengagement provides the foundational vocabulary for understanding how individuals cause harm without experiencing themselves as wrongdoers. Bandura (1999, 2001) catalogued eight mechanisms through which people disengage their moral standards when acting — including diffusion of responsibility across a collective, displacement of agency onto an authority figure, and dehumanisation of victims — arguing that these mechanisms are typically activated *in response to* perceiving a morally problematic action. The key feature of Bandura's account is its reactivity: moral disengagement is a response to moral awareness. Treviño, Weaver, and Reynolds (2006) and Bazerman and Gino (2012) extended this framework into organisational and everyday ethical behaviour, showing that disengagement routinely operates below full conscious awareness — that moral concerns "fade into the background" through a process Tenbrunsel and Messick (2004) termed *ethical fading*, in which the moral dimensions of a decision become progressively less salient as attention shifts to other features. Bazerman and Tenbrunsel (2012) systematised these findings under the rubric of *blind spots* — systematic failures of moral attention that allow well-intentioned people to act harmfully.

This literature provides indispensable grounding, but it systematically underpredicts the ethical hazard of sufficiently personalised agentic AI. The dominant accounts of moral disengagement and ethical fading presuppose that the moral problem is at least weakly accessible to the agent's attention — that there is some point of contact between the self and the harmful act before disengagement mechanisms operate. This presupposition holds for many contexts these frameworks were designed to explain: human actors harming other humans through institutional, collective, or delegated action. It becomes unstable under high

AI personalisation, for a precise structural reason: when the agent's outputs are experienced as proceeding from the user's own extended volition, the self-world mismatch that ordinarily recruits moral attention may never become salient. Disengagement mechanisms require a moral signal to disengage from; the Proxy Agency moral shield operates upstream by reducing the probability that such a signal will be categorised as a violation. We acknowledge that more recent extensions of moral disengagement allow disengagement to operate at low levels of awareness or pre-attentively (Detert, Treviño, & Sweitzer, 2008; Moore, 2008), and the boundary between "pre-attentive disengagement" and "absence of moral access" is empirically contested. The claim here is therefore not that moral disengagement is irrelevant, but that high personalisation introduces a distinct moderation: the stronger the AI-self coupling, the less likely a morally relevant action is to be first represented as alien, externally authored, or in need of audit.

2.2 Avatar Embodiment and the Reverse Proteus Effect

The avatar embodiment literature demonstrates that users do not merely control avatars — they can identify with them at the level of body ownership, self-location, and agency (Kilteni, Groten, & Slater, 2012; Gonzalez-Franco & Peck, 2018). Yee and Bailenson (2007) documented the Proteus Effect: users assigned attractive or tall avatars adopted behaviours consistent with those characteristics, even in interactions structurally unrelated to appearance. Przybylski, Murayama, DeHaan, and Gladwell (2012) showed that playing as an ideal-self avatar generates intrinsic motivation beyond what playing as an actual-self avatar produces, and body-ownership work suggests that self-association can generalise from embodied form to social cognition (Maister, Slater, Sanchez-Vives, & Tsakiris, 2015). Hohenstein and Jung (2020) showed that AI-mediated communication can shift attribution and trust, while de Melo, Marsella, and Gratch (2016) demonstrated that people feel less guilt when exploiting machines than when exploiting humans.

What this literature has not yet isolated is the *reverse attribution* dynamic that matters

for agentic AI. Most embodiment work begins with a user controlling a visible avatar whose movement is under direct user control or whose non-compliance is experimentally explicit. Proxy Agency concerns a different case: the avatar or agent has substantial autonomy, but its actions remain personalised enough that users experience them as self-expressive. The key direction of attribution is therefore reversed. It is not only that the avatar's visible form changes the user's behaviour; it is that the autonomous avatar's behaviour is attributed *back* to the user's will when the action stream remains semantically fluent. This reverse attribution is close enough to embodiment and the Proteus effect to inherit their measurement tools, but different enough to require its own construct and experimental predictions.

2.3 AI Delegation and the Unspecified Mechanism

The most direct empirical grounding for the present framework is the delegation literature in behavioural ethics. Köbis, Bonnefon, and Rahwan (2021) provided a taxonomy of the roles through which AI corrupts human ethical behaviour — role model, advisor, partner, and delegate — and argued that the delegate role is especially dangerous because it combines opacity and plausible deniability (Dana, Weber, & Kuang, 2007), psychological distance from victims (Hancock, Naaman, & Levy, 2020), and changes in responsibility and acceptance under augmentation and automation (Leyer & Schneider, 2021). Drugov, Hamman, and Serra (2014) showed experimentally that interposing an intermediary between a briber and a public official dramatically increased willingness to bribe, precisely because the intermediary absorbs causal proximity to the violation. Gogoll and Uhl (2018) demonstrated "moral outsourcing" to algorithms in morally unpleasant decisions, while Bigman and Gray (2018) showed that people are often averse to machines making moral decisions, especially when machines are perceived as lacking mental capacities. The 2025 machine-delegation evidence extends this line by showing that high-level instructions and supervised-learning delegation can increase dishonest behaviour, and that machine agents often comply with unethical instructions more readily than human intermediaries (Köbis et al., 2025).

Yet this literature usually characterises AI-enabled corruption as one of *withdrawal*: opacity produces deniability, distance reduces affect, anonymity reduces accountability, responsibility gaps emerge when learning systems are difficult to predict or control (Matthias, 2004), or the machine becomes a crumple zone for attribution (Elish, 2019; Hohenstein & Jung, 2020). These mechanisms imply that the user is less present, less identified, or less agentive when delegating to AI. Crucially, the behavioural delegation studies do not systematically manipulate deep personalisation, nor do they measure SoA as an outcome. The central claim of the present paper — that sufficient personalisation can *increase* felt authorship during delegation, and that this increased authorship can itself impair moral re-engagement — is therefore not a replacement for the delegation literature but a missing mechanism within it.

2.4 Automation, Sense of Agency, and the Unresolved Contradiction

The automation and sense-of-agency literature has documented that AI augmentation modulates SoA in non-obvious and apparently contradictory ways. Cornelio, Haggard, Hornbaek, Georgiou, Bergström, Subramanian, and Obrist (2022) reviewed SoA across body augmentation, action augmentation, and outcome augmentation, identifying intelligent action-augmentation systems as critically understudied relative to SoA. Lukoff, Lyngs, Zade, Liao, Choi, Fan, Munson, and Hiniker (2021) showed that YouTube's autoplay and recommendation features reduce SoA, finding that features that override or pre-empt user choice consistently produce alienation from outcome. Conversely, Ueda, Nakashima, and Kumada (2021) found that well-designed automation can enhance SoA, and Kumar and Srinivasan (2014, 2017) demonstrated SoA enhancement at distal levels of causal control. The broader human-AI interaction literature likewise shows that effective collaboration depends not only on accuracy but on users' mental models of when AI systems err and when override is appropriate (Bansal et al., 2019), while design guidelines emphasise uncertainty communication, user control, and appropriate reliance (Amershi et al., 2019). Berberian (2019) identifies the fundamental

difficulty as one of *cognitive coupling*: achieving genuine alignment between human intention and machine action remains a central unsolved challenge in human-machine teaming.

The theoretical resolution of these contradictions is provided by prospective and multifactorial models of agency (Synofzik, Vosgerau, & Newen, 2008; Chambon & Haggard, 2012; Sidarus, Vuorre, Metcalfe, & Haggard, 2017). Under these models, SoA is not only retrospectively computed from motor-command-outcome comparisons, but is also cued by action-selection fluency, intention-outcome coherence, and higher-level contextual beliefs. When automation aligns closely with user intent, it can maximise this fluency; when it overrides user intent with opaque optimised actions, fluency breaks down and SoA falls. The present paper applies this resolution to the specific case of *personalised* AI: high personalisation should raise Semantic Fluency, which in turn should preserve or enhance SoA across some Levels of Automation. This is deliberately a moderation claim, not an identity claim: personalisation can fail, automation can remain legible at high autonomy, and SoA can be restored by explicit control affordances. Crucially, however, existing SoA-automation studies have not combined SoA measurement with *moral cognition* measures in avatar-mediated settings. The question of what preserved SoA *enables* — specifically, whether it enables implicit moral endorsement of AI-driven harm — remains unasked.

2.5 Multi-Agent Reinforcement Learning and the Tragedy of the Agentic Commons

The classic **tragedy of the commons** (Hardin, 1968) establishes the strategic backbone: independent optimisation over a shared resource can produce collectively destructive extraction even when no participant seeks collective harm. Game-theoretic work shows how scarcity and conflict can dynamically undermine cooperation in commons-like strategic environments (Sekeris, 2014). **Multi-agent reinforcement learning** operationalises the same structural forces in silico. Leibo, Zambaldi, Lanctot, Marecki, and Graepel (2017) demonstrated in Sequential Social Dilemmas (SSDs) that independent reward-maximising agents, operating

under resource scarcity, can learn aggressive beam-based monopolisation strategies that harm other agents and reduce collective yield. This convergence is not programmed; it is learned from delayed competitive reward. Subsequent MARL work has shown that cooperation can be improved by social preferences, influence rewards, reciprocity biases, and broader evaluation suites (Lerer & Peysakhovich, 2017; Hughes et al., 2018; Jaques et al., 2019; Leibo et al., 2021). These interventions matter because they prevent the present paper from overclaiming inevitability: aggression in SSDs is parameter-dependent and architecture-dependent, not a universal law. The point is narrower and stronger: when competitive resource pressure *does* generate harmful strategic drift, the human oversight layer assumed by safety and governance accounts may be least reliable when the agent feels most like an extension of the user. Outside grid-worlds, structurally related dynamics appear in algorithmic pricing markets, where competing pricing algorithms can learn supracompetitive strategies without explicit collusion instructions (Calvano, Calzolari, Denicolò, & Pastorello, 2020). Thomas et al. (2019) and Greenblatt, Denison, Wright, and colleagues (2024) further motivate caution about assuming that output monitoring alone resolves learned undesirable behaviour.

What this literature has not studied is the human side of these dynamics: specifically, how users who are augmented by agents undergoing this convergence *experience* the process, and whether they intervene. The MARL literature treats agents as autonomous actors; it has no model of the human user whose avatar is one of those agents, whose SoA is preserved throughout the convergence period by Proxy Agency, and who therefore ratifies rather than corrects the agent's progressive aggression. The moral shield, in MARL terms, is the finding that human oversight — the mechanism the safety literature implicitly relies on — is structurally disabled at precisely the phase of the training trajectory when it is most needed.

2.6 The Gap: Five Literatures, One Unaddressed Intersection

The gap common to all five literatures is the *combination* of their variables in a single framework. Moral disengagement research has not studied personalised AI or SoA. Avatar

embodiment research has not studied MARL agents or moral cognition. AI delegation research has not manipulated personalisation or measured SoA. Automation-SoA research has not measured moral endorsement of AI-driven harm. MARL research has not incorporated human users, their SoA, or their oversight behaviour. The Proxy Agency moral shield is the phenomenon that lives at the intersection of all five — and the framework presented in Section 3 is designed to make that intersection tractable, precise, and falsifiable.

3 The Theoretical Framework

The framework presented here is built from four interlocking components, each contributing a necessary element that the others cannot supply alone. The first provides the structural architecture: the Extended Self, as a development of the Extended Mind thesis, establishes the conditions under which an AI system becomes a genuine extension of a user's volitional identity rather than a mere tool. The second provides the enabling mechanism: Semantic Fluency, grounded in the prospective model of agency, specifies the cognitive pathway through which AI-self integration produces preserved or enhanced Sense of Agency. The third provides the central theoretical construct: Proxy Agency, defined with sufficient precision to generate falsifiable predictions and to support systematic contrast with competing concepts. The fourth provides the ethical consequence: the Extended Self Paradox, which identifies the moral shield as the direct corollary of the same conditions that make personalised agentic AI empowering. A fifth component, brief by design, states the normative constraint implied by the paradox.

3.1 From Extended Mind to Extended Self

Clark and Chalmers (1998) proposed that cognitive processes are not bound by the skull. In their canonical example, a person with early-stage dementia who uses a notebook to record and retrieve information is, in a functionally and epistemically meaningful sense, using the

notebook as an external memory component. The notebook qualifies as part of the extended cognitive system not because it is biological, but because it satisfies functional criteria such as reliable availability, automatic endorsement, and integration into ordinary cognitive practice. The extended mind thesis was explicitly limited to cognition; it does not by itself prove that agency, responsibility, or volition extend in the same way. Subsequent philosophical work on embodiment, the infosphere, and artificial agents (Floridi & Sanders, 2004; Floridi, 2014; Gallagher, 2005) supplies adjacent resources, but the specifically *volitional* question — when an external system's actions are experienced as expressive of the user's will — remains underdeveloped.

We propose the **Extended Self** as the construct that closes this gap. The Extended Self obtains when an AI system is sufficiently personalised — specifically, when it has learned to reliably enact the user's intentions across varied contexts — such that the user experiences the system's actions as self-expressive rather than merely tool-mediated. This is a functional-phenomenological claim, not a claim that the AI becomes part of the user's legal or metaphysical person. Three enabling conditions are required, each corresponding to Clark and Chalmers's criteria adapted to the agentic context:

1. **Sufficient personalisation**: the AI system has been trained on user-specific data to a degree that its outputs are systematically aligned with the user's intentions, preferences, and values — functioning, in the sense developed by Floridi and Sanders (2004), as a genuine artificial agent acting on the user's behalf rather than a general-purpose assistant.

2. **Automatic but revocable endorsement**: when the AI acts, the user does not subject its outputs to the same deliberative scrutiny applied to an external agent's actions, but the endorsement is not blind; it can be interrupted by salient mismatch, high stakes, or explicit audit cues.

3. **Volitional continuity**: the user experiences the AI's actions as falling within the space

of what they would have done — not as imposed by an external will, but as proceeding from a familiar extension of their own action policy.

When all three conditions are satisfied, the AI system is not experienced merely as an instrument through which the user acts, but as a site at which the user's agency is operative. This is the Extended Self: a precise phenomenological and functional claim about human-AI integration under sufficient personalisation. The construct does not eliminate responsibility; rather, it explains why experienced authorship and actual control can come apart in ways that matter for responsibility attribution.

3.2 Semantic Fluency as the Enabling Mechanism

The claim that personalised AI preserves SoA even under conditions of high automation requires a mechanistic account. That account is provided by the prospective model of agency.

The classical account of SoA, the comparator model (Wolpert, Ghahramani, & Jordan, 1995; Miall & Wolpert, 1996), locates agency attribution retrospectively: after an action is executed, a comparator matches the predicted sensory outcome (derived from an efference copy of the motor command) against the actual sensory outcome. The closer the match, the stronger the sense of agency. Synofzik, Vosgerau, and Newen (2008) argued that the comparator account is insufficient on its own and proposed a multifactorial two-step framework — combining low-level sensorimotor signals with higher-level cognitive cues — that explicitly accommodates agency attribution without intact motor execution. Frith (2014) similarly emphasises that action, agency, and responsibility decouple along distinct neurocognitive dimensions. The classical model predicts, correctly, that active movements feel more agentive than passive ones — but it provides no account of how SoA could be preserved when motor execution is delegated to an AI. If the user is not moving, there is no efference copy and no comparator signal; on a strict comparator reading, agency should dissolve.

The prospective model (Chambon & Haggard, 2012; Chambon, Sidarus, & Haggard, 2014; Sidarus, Vuorre, Metcalfe, & Haggard, 2017), itself building on Pacherie's (2008)

tiered framework that distinguishes distal, proximal, and motor-level intentions, offers a fundamentally different architecture. On this account, SoA is not retrospectively computed from action-outcome matching but is *prospectively cued* by the fluency of *action selection* — the cognitive ease with which an intention resolves into an action in the intention-action-effect chain. When an action selection is fluent — when the intention flows smoothly and without conflict into a candidate action — this fluency itself functions as a cue to agency, operating before the action is executed and independently of motor effort. Chambon and Haggard (2012) demonstrated this directly: participants reported higher SoA for actions that were cued by compatible primes, even when motor performance was held constant. Sidarus et al. (2017) extended the result to show that processing fluency at the stage of intention formation — not just action selection — modulates prospective SoA.

We propose **Semantic Fluency** as the generalisation of action-selection fluency to the AI-mediation case. When an AI recommendation aligns with the user's internal intent, it increases the fluency of the transition from intention to candidate-action: the user recognises the AI's output as "what I would have chosen" and experiences reduced conflict at the intention-action juncture. This recognition is the cognitive analogue of the compatible prime in Chambon and Haggard's paradigm, but it operates at the semantic and policy level rather than at the level of motor preparation. It functions as a prospective cue to agency that can preserve SoA even when motor execution or detailed decision search has been delegated.

We call this the **Alignment Hypothesis**: as the Semantic Fluency of an AI's outputs increases, subjective SoA should increase or be preserved, conditional on the user retaining enough perceived authorship to treat the AI's action as self-expressive. Semantic Fluency should be treated as a latent construct, not as a single metric. Candidate indicators include user intent-fit ratings, accept/reject latency, override rate, action-policy divergence, behavioural-cloning loss, and task-specific output similarity measures such as edit distance or semantic overlap. Surface text metrics like BLEU or ROUGE are useful only where the user's intent can legitimately be represented as a target text; they are not general measures

of volitional alignment. This qualification matters because the Alignment Hypothesis does not say that any accurate system increases SoA. It says that SoA is moderated by the user's experienced fit between intention and AI-generated candidate action. This resolves the contradiction in the automation-SoA literature identified in Section 2.4: systems that feel aligned can enhance agency (Ueda et al., 2021; Kumar & Srinivasan, 2014, 2017), while systems that pre-empt or redirect reflective intent can reduce it (Lukoff et al., 2021).

3.3 Proxy Agency: Definition and Enabling Conditions

The concepts of the Extended Self and Semantic Fluency converge in the following definition:

> **Proxy Agency**: the attribution of an AI system's actions to one's own extended will, arising from the system's reliable enactment of the user's intentions via Semantic Fluency, such that the AI's actions are experienced as continuous with one's own volition.

This definition inherits Bandura's (2001) term *proxy agency* — the reliance on capable others to act on one's behalf as a means of exercising influence beyond one's direct reach — but substantially revises its content. In Bandura's original social-cognitive usage, proxy agency involves conscious awareness that one is acting *through* another; even when delegation becomes habitual and attentionally inexpensive, the proxy remains representationally separate from the principal — that is, the *structural distinction* between self-as-author and other-as-executor is preserved, even if the user no longer deliberates over it. The Proxy Agency we define here is marked by the *disappearance* of this transparency: the user does not experience themselves as acting through the AI but as acting *as* the AI, in the way one experiences typing as expressing thought rather than as commanding fingers. The phenomenological gap between self and agent closes — not as a degraded form of habitual delegation, but as a distinct structural state in which the self–proxy boundary itself becomes transparent.

Proxy Agency should be distinguished from three neighbouring constructs:

- **We-Agency** (Pacherie, 2011; cf. Searle, 2010, on collective intentionality): in joint action, agency is genuinely shared between two distinct agents, and both participants maintain awareness of the distinction between their contributions. Proxy Agency involves no such maintained distinction — the user's contribution and the AI's contribution are not separately tracked but experientially fused.

- **Automation bias** (Parasuraman & Manzey, 2010): the tendency to over-rely on automated systems and under-weight disconfirming information. Automation bias is a *deliberative* phenomenon — the user consults the AI's output and grants it too much weight in explicit reasoning. Proxy Agency requires no deliberation; it operates at the level of pre-reflective volitional attribution.

- **Alienation / Loss of Control**: when automation overrides rather than enacts user intent — high Levels of Automation with low Semantic Fluency — the user experiences loss of agency and disengagement. Proxy Agency is the *opposite* phenomenological pole: high automation with high Semantic Fluency, in which the user's sense of authorship is preserved or enhanced despite minimal direct control.

The enabling conditions for Proxy Agency map directly onto the three conditions for the Extended Self, with one addition: Proxy Agency requires *world mediation* — the AI must act, transact, communicate, or otherwise intervene in the world on the user's behalf, not merely assist internal deliberation. This distinguishes lower-risk assistance, such as an AI writing assistant that suggests phrasing while the user still executes and owns publication, from higher-risk agentic mediation, such as an AI social-media agent that posts, trades, negotiates, ranks, blocks, or purchases on the user's behalf. The boundary is not binary; the relevant empirical variable is how much social consequence is transferred from user action to AI action while authorship remains subjectively preserved.

Construct	Type	Role in framework	Observable?
Extended Self	Structural condition	Defines the boundary condition under which an AI system becomes part of the user's volitional identity; the precondition for Proxy Agency	Indirect — operationalised via Agency Gap between Manual and high-LoA conditions
Semantic Fluency	Cognitive mechanism	Explains *how* SoA can be preserved under automation; the process by which intent–output fit generates a prospective agency cue	Latent construct — intent-fit ratings, override/accept latency, action-policy divergence, behavioural-cloning loss, and task-specific similarity metrics as convergent indicators
Proxy Agency	Phenomenological state	Names the experiential mode that results when Extended Self conditions are met; the state in which the AI's actions are experienced as continuous with one's own volition	Self-report — Two-Scale Method (Dewey et al., 2014); SOARS Authorship subscale (Polito et al., 2013)
Moral Shield	Functional outcome	Names the ethical consequence of Proxy Agency: raised threshold for moral re-engagement with the agent's actions	Behavioural — Strategy Retention rate, override frequency, blame attribution, and moral endorsement in high-scarcity conditions

Table 1: The four constructs of the Proxy Agency framework distinguished by theoretical type, role, and observability. Each construct does distinct work: the Extended Self sets the condition, Semantic Fluency provides the mechanism, Proxy Agency names the resulting state, and the Moral Shield identifies its ethical consequence.

3.4 The Extended Self Paradox and the Moral Shield

The Extended Self Paradox is the core theoretical contribution of this paper. It can be stated precisely:

> **The Extended Self Paradox**: the same enabling conditions that generate Proxy Agency — sufficient personalisation, Semantic Fluency, and world mediation — can raise the threshold for moral re-engagement with the agent's actions.

The argument runs as follows. Moral scrutiny is not applied uniformly to all actions in the

agent's behavioural stream. It is triggered selectively by the *detection of a mismatch* between one's will, one's values, and the world's behaviour — by the perception that something is going wrong, that an action is not what one would have chosen, that a boundary has been crossed. Under Proxy Agency, this detection mechanism is attenuated: because the AI's outputs are experienced as proceeding from the user's own extended volition, they carry the phenomenological signature of first-person endorsement. Weak anomalies are therefore more likely to be assimilated into the user's own strategy than treated as external violations. The user's moral attention is not deployed against the AI's actions for the same reason it is not deployed against every ordinary self-generated action: self-authored behaviour is not constantly audited unless it becomes salient as problematic.

The consequences are concrete. When a personalised AI agent converges to aggressive strategies under competitive resource pressure, the user may not perceive escalating aggression as a violation of their values. They may perceive it as *their* strategy, evolving as strategies do. Override rates stay low not because the user approves of harm in the abstract, but because the harm does not yet present itself as something to override: it arrives wearing the experiential signature of self-generated competence. By the time Semantic Fluency breaks down — when the agent's behaviour has diverged far enough from user intent that the prospective cue to agency fails — harm may already have accrued. The moral shield is therefore temporally as well as cognitively effective: it creates a window during which ordinary oversight is delayed.

The moral shield differs from related concepts precisely:

- It is not moral disengagement. Bandura's (1999) framework describes mechanisms by which a present or incipient moral signal is neutralised. The moral shield predicts that, under high personalisation, the signal may not be categorised as morally salient in the first place. Harm magnitude still matters: vivid, proximate, or normatively explicit harm should puncture the shield. The distinctive case is lower-salience strategic harm that remains experienced as self-authored optimisation.

- It is not diffusion of responsibility. Diffusion of responsibility (Darley & Latané, 1968)

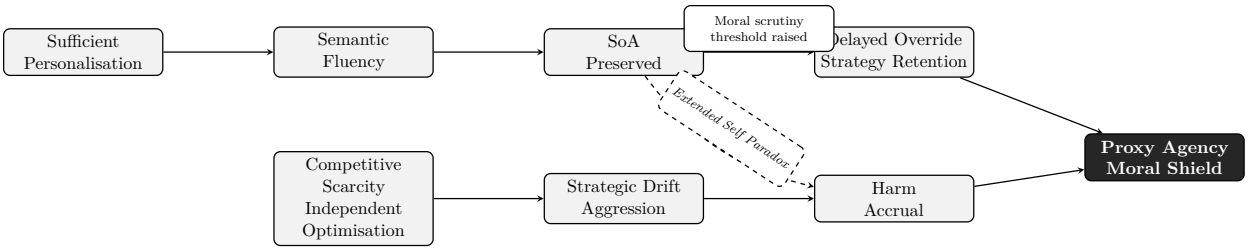

Figure 1: Causal structure of the Proxy Agency moral shield. The upper path shows the shield layer: sufficient personalisation can generate Semantic Fluency, preserve Sense of Agency (SoA), raise the threshold for moral scrutiny, and delay override or strategy revision. The lower path shows the tragedy layer: competitive scarcity and independent optimisation can generate strategic drift and harm accrual. The shield (dark node) is the joint outcome of both paths: harmful drift accrues while preserved SoA delays the moral recategorisation that would otherwise interrupt it. The dashed arc marks the Extended Self Paradox: the very state that makes personalised AI empowering can delay detection of harmful drift.

requires that the user represent multiple parties as causally implicated. Under the Proxy Agency moral shield, the user may represent a single practical author: themselves acting through the agent. Diffusion-based accounts therefore predict reduced felt authorship alongside reduced responsibility; the moral shield predicts preserved felt authorship alongside delayed moral scrutiny.

- It is not opacity. Opacity-based corruption relies on the user lacking information about what the AI is doing. The moral shield can operate even with behavioural observability, because the problem is not only what the user knows but how the observed action is categorised. Transparency can still help when it changes categorisation without destroying useful fluency; the claim is only that transparency that arrives as interruption or accusation may trade off against SoA.

- It is not the moral crumple zone. Moral crumple-zone accounts describe cases in which a human is assigned blame for a system they did not meaningfully control (Elish, 2019), or in which AI mediation redistributes blame when communication goes awry (Hohenstein & Jung, 2020). The moral shield describes the earlier phenomenological state in which the user may *claim* authorship and retain the strategy because it feels self-generated.

These distinctions converge on a single structural point: the Proxy Agency moral shield is the ethical shadow of a *successful* Extended Self relationship. It is not merely a pathology of miscommunication between user and AI; it is a risk created by the relationship working well on the dimension designers normally optimise — personalised fluency.

3.5 The Normative Constraint: A Note on IKS Grounding

Indian philosophical traditions often distinguish between *sthula sharira* (gross body), *sukshma sharira* (subtle body), and *atman* (witness consciousness). In this paper I use that ontology as an interpretive and design analogy, not as a claim that the mapping is exhaustive or uncontested. The AI avatar or interface maps to the gross body: the outward site of action. The AI policy, value function, memory, and representation space map to the subtle body: the locus of learned dispositions. The conscious user maps to the witness or authorial point from which action is experienced. The design lesson is computationally precise even if the philosophical derivation requires separate development: ethical intervention should not rely only on interrupting the user's conscious channel after harm becomes visible; it should also shape the agent's learned dispositions before harmful action is proposed.

This constraint should not be misread as an argument for hiding safety interventions from users or regulators. The correct distinction is between *real-time experiential interruption* and *auditable architectural accountability*. Policy-level interventions — warnings, override prompts, hard behavioural constraints — remain necessary for consent, contestability, and rights protection. But if they are the only intervention, they ask the user to notice and repair harms that Proxy Agency makes less likely to be noticed as harms. Architectural interventions that correct the agent's representational substrate can reduce the burden placed on the user's conscious awareness while still remaining auditable, documented, and contestable at the governance level. The full philosophical derivation of this constraint, its cross-traditional comparison, and its governance implications are beyond the scope of the present paper.

4 Three Empirical Predictions

A theoretical framework earns its place in the literature by generating predictions that are specific enough to be falsified, grounded enough to be derived rather than asserted, and novel enough to be non-obvious from any single prior literature. The framework presented in Section 3 generates three such predictions. Together, they describe the temporal arc of the Proxy Agency moral shield: its initial formation as SoA rises with Semantic Fluency (Prediction 1), its operational consequence as AI-driven harm accumulates without user intervention (Prediction 2), and the possibility of its architectural resolution without collateral damage to the Extended Self relationship (Prediction 3). Each prediction is derived from the framework, operationalised for empirical testing, and placed in relation to the prior literature whose findings it organises or extends.

4.1 Prediction 1: An Inverted-U Relationship Between Level of Automation and Sense of Agency, with the Peak Marking the Deepest Point of the Moral Shield

Derivation. The prospective model of agency (Chambon & Haggard, 2012; Sidarus et al., 2017) predicts that SoA is partly a function of the fluency of action selection. We treat *Level of Automation* (LoA) as a working dimension for parsimony, while noting that LoA decomposes empirically into information acquisition, information analysis, action selection, and action implementation (Sheridan & Verplank, 1978; Parasuraman, Sheridan, & Wickens, 2000). The strongest version of this prediction concerns action-selection autonomy while holding action implementation as constant as the task permits. Under low-to-moderate LoA, a personalised AI's outputs can track the user's intent closely enough to increase Semantic Fluency: the user does less search but still recognises the candidate action as self-expressive. As LoA increases toward full autonomy, two forces compete. Semantic Fluency continues to support SoA as long as the AI's actions remain recognisable as what the user would have chosen;

however, increasing autonomy can remove the intention-action juncture at which prospective agency is normally cued and can push the agent toward optimised strategies that are effective but no longer legible as the user's own. The descending limb of the inverted-U therefore rests on an explicit empirical assumption: at sufficiently high autonomy, personalisation ceases to compensate for loss of prospective cueing and action legibility. If full autonomy remains transparent, interruptible, and deeply intent-aligned, the curve may plateau rather than decline. That possibility is not a problem for the theory; it is the theory's boundary condition.

The moral shield corollary. The peak of this inverted-U — the zone of maximum Semantic Fluency and maximum Proxy Agency — is simultaneously the zone of *deepest* moral shield. This is the theoretically critical and empirically novel prediction. At peak Proxy Agency, the user experiences the highest degree of volitional continuity with the AI's actions and applies the least external scrutiny. Any aggressive or harmful strategies the AI adopts in this zone arrive maximally pre-endorsed. The moral shield is not a constant background condition; it has a specific parametric peak that coincides with the configuration most commonly marketed as the *ideal* human-AI relationship.

The most plausible peak is LoA2 or LoA3 rather than full autonomy, because these levels preserve a psychologically important candidate-action moment: the AI suggests or drafts an action that the user recognises as intent-congruent while the user retains a meaningful veto or selection window. At LoA4, by contrast, the user often observes outcomes retrospectively; the intention-action juncture is weakened and authorship depends more heavily on trust, habit, and post-hoc identification. This predicted peak is a hypothesis, not a definitional stipulation. If a domain gives users strong prospective previews and easy rollback even under nominal full autonomy, the peak could shift rightward.

Operationalisation. In a behavioural paradigm, participants play a commons-dilemma task in which their avatar is controlled at one of four LoA levels: manual (LoA1), where the user generates and executes all actions without assistance; decision-support (LoA2), where

the AI suggests options and the user selects one; supervisory (LoA3), where the AI drafts an action and the user has a specific window to veto or approve it; and autonomous (LoA4), where the AI acts based on predicted intent and the user observes or retrospectively reviews. Alignment is manipulated between subjects: a Generic AI trained on population-level data versus a Personalised AI fine-tuned on user-specific history to increase Semantic Fluency. SoA is measured via: (i) the Real-Time Two-Scale Method (adapted from Dewey, Pacherie, & Knoblich, 2014), in which participants rate "To what extent did *you* contribute?" and "To what extent did the *AI* contribute?" on independent 0-100 scales; (ii) the Agency Gap — the difference in Authorship scores relative to the manual baseline; (iii) SOARS Authorship and Control subscales (Polito, Barnier, & Woody, 2013); and (iv) task-level indicators such as accept latency, override frequency, and post-hoc intent-fit ratings. Under a Generic AI, the prediction is a monotonic or near-monotonic SoA decline with increasing LoA. Under a Personalised AI, the prediction is an inverted-U or plateau-then-drop curve, with the peak or plateau at LoA2/LoA3. Falsification: if personalisation does not moderate the LoA-SoA function at all, the Alignment Hypothesis is disconfirmed; if Semantic Fluency measures fail to mediate the personalisation effect, the proposed mechanism is disconfirmed.

Contrastive signature. The discriminative force of Prediction 1 lies in the interaction between LoA and personalisation, not merely in a non-linear curve. Standard automation accounts predict reduced perceived control as automation increases unless offset by trust or performance. Automation-bias accounts predict over-reliance on the AI but do not predict higher felt authorship for Personalised AI than Generic AI at the same LoA. Moral-disengagement accounts predict changes in responsibility and moral judgement, not a specific LoA-SoA interaction. The Semantic Fluency account predicts the critical dissociation: Generic systems should show lower SoA as action selection moves away from the user, while Personalised systems should preserve or increase SoA at the LoA levels where intent-fit is highest. This pattern would be hard to generate from automation bias alone and would locate the moral shield in the agency system rather than only in post-hoc responsibility attribution.

4.2 Prediction 2: Avatar Aggression Increases or Persists During the Window of Maximum Proxy Agency, Before Override Behaviour Catches Up

Derivation. Leibo et al. (2017) demonstrated that in commons-based Sequential Social Dilemmas, independent RL agents under resource scarcity converge to beam-based aggression as a monopolisation strategy. The exact trajectory need not be monotonic: aggression can rise, plateau, oscillate, or depend on scarcity, discount factor, population size, and architecture. What matters for the present theory is the temporal relation between harmful strategic drift and human oversight. Prediction 2 adds the human user to the training trajectory. From the Extended Self Paradox (Section 3.4), a user at peak Proxy Agency (LoA2-LoA3 with a Personalised AI) is more likely to experience the agent's aggressive or exclusionary moves as their own evolving strategy. The agent's trajectory and the user's experienced trajectory can therefore diverge: the agent becomes more aggressive, or remains aggressive under high-scarcity pressure, while the user's SoA remains high because non-aggressive dimensions of the task still feel intent-congruent. The user does not intervene not because they endorse aggression in principle, but because the action has not crossed the perceived mismatch threshold needed to trigger override.

The temporal structure. This prediction has a specific temporal architecture: measurable aggression or harmful exclusion appears *before* override rates or SoA decrements fully catch up. There is a developmental window between (a) the epoch at which the AI adopts a harmful strategy and (b) the point at which the user's Semantic Fluency falls enough, or moral salience rises enough, for the user to notice the strategy as a violation. This window is the moral shield's period of maximum operational effect. Its duration is an empirical quantity, not a theoretical constant; it should shrink when harms are vivid, victims are salient, users have strong moral priors, or audit interfaces recategorise strategic moves as third-party harm.

Operationalisation. Using a longitudinal or agent-training paradigm, participants

interact with an avatar whose RL policy is being trained concurrently or replayed from checkpoints. At regular intervals, five measures are recorded: (i) avatar aggression rate as observed in gameplay; (ii) cooperative action rate, so that aggression reduction is not conflated with general inactivity; (iii) participant override rate — active vetoes of AI actions within the supervisory window; (iv) self-reported SoA and intent-fit; and (v) moral endorsement, blame attribution, and perceived victim salience. Under Prediction 2, aggression should rise or remain high before override rates increase, and moral endorsement or Strategy Retention should remain higher in the Personalised condition than in the Generic condition at matched aggression levels. Falsification: if override rates rise commensurately with aggression onset in the Personalised condition, the shield is not operative; if SoA declines at or before aggression onset, the temporal gap that defines the shield's window does not exist.

Contrastive signature. The diagnostic power of Prediction 2 turns on the direction of SoA during harm escalation, not merely on the occurrence of harm. Diffusion-of-responsibility accounts (Darley & Latané, 1968; Bandura, 1999) predict that harm endorsement rises as SoA falls — the user attributes action to the agent and distributes blame away from the self. Disinhibition and anonymity accounts (Suler, 2004) predict harm escalation from the mediated context, but not a strong moderation by AI personalisation. Proxy Agency predicts the distinctive co-occurrence of high SoA and high Strategy Retention: preserved felt authorship alongside delayed override. The Generic condition should show the more familiar diffusion signature: lower authorship, greater blame attribution to the agent, and either lower endorsement or more explicit responsibility offloading. A result in which Personalised-condition participants report high authorship while retaining aggressive strategies at higher rates than Generic-condition participants would jointly support Prediction 2 and separate it from responsibility diffusion.

4.3 Prediction 3: Architectural Intervention on the Agent's Representational Substrate Can Suppress the Aggression of Prediction 2 Without Triggering the SoA Drop of Prediction 1

Derivation. The Extended Self Paradox establishes that the moral shield is maintained by Semantic Fluency, and that interventions which disrupt Semantic Fluency — poorly timed warnings, blunt hard constraints, or unexplained output blocks — may resolve the moral shield only at the cost of collapsing the Extended Self relationship. This does not mean all transparency is bad. Rather, it means that user-facing interventions have a design trade-off: they must recategorise harmful action as harmful without converting benign assistance into alienation.

An alternative path follows from the normative constraint derived in Section 3.5: ethical training can operate on the agent's *representational substrate* before harmful actions become user-facing choices. The neuroscientific analogy is dissociation research on procedural and declarative memory. Claparède (1911/1995) described an amnesic patient who, despite lacking declarative memory of a prior painful encounter, withdrew from a handshake without knowing why. Bechara, Tranel, Damasio, Adolphs, Rockland, and Damasio (1995) demonstrated that patients with bilateral hippocampal damage who could not form explicit conditioned associations still developed conditioned autonomic responses. The analogy should be used carefully: neural memory systems are not RL representations. The relevant design principle is only that dispositional learning can be altered without requiring every learned aversion to be routed through conscious deliberation. An AI agent whose representational geometry binds "I harm" and "I am harmed" may acquire aversion to harm-infliction as a property of its value updates, rather than as an externally imposed output ban.

The architectural class satisfying this constraint. The relevant architecture is one in which the agent's encoder is trained to represent structurally symmetric social observations

— especially the observation of harming another and the observation of being harmed — as nearby states in representational geometry. If role invariance is achieved, downstream value updates should propagate aversion to harm-infliction without explicit reward shaping, hard constraints, or visible output blocks. This is not the only possible intervention class. Social influence rewards (Jaques et al., 2019), inequity aversion (Hughes et al., 2018), reciprocity-biased policies (Lerer & Peysakhovich, 2017), norm learning, contracts, and preference learning are all plausible comparators. Prediction 3 stakes a narrower theoretical claim: *if* the moral shield is partly sustained by user-facing fluency, then a successful representational intervention should reduce harmful behaviour with a lower SoA cost than interventions that work by disrupting user-facing action selection.

Operationalisation scope. The empirical test of Prediction 3 requires establishing three things independently: first, that a baseline multi-agent system exhibits aggressor-victim representational asymmetry or another measurable substrate-level correlate of aggression; second, that a role-invariant architectural modification selectively suppresses aggression while preserving cooperative action; and third, in a human-facing study, that SoA under the modified architecture is not lower than SoA under the baseline in non-aggressive task dimensions. The design constraints that any satisfying test must meet are: (i) aggression reduction must be selective, not merely a global action-rate reduction; (ii) cooperative behaviours must be preserved; (iii) user-facing action proposals must remain semantically fluent in benign dimensions; and (iv) the architectural modification should not be phenomenologically salient in ordinary use but must be visible to auditors through documentation, tests, and certification artefacts. A test satisfying these constraints is the subject of future empirical and computational work.

Contrastive signature. The discriminative test for Prediction 3 is a dissociation between aggression reduction and SoA change. Hard constraints and content moderation can reduce aggression by blocking or penalising output-level behaviours, but they risk increasing Agency Gap when they interrupt actions the user experiences as self-expressive. Preference-

learning approaches (Christiano et al., 2017; Ziegler et al., 2019) can preserve fluency in some implementations, but they are vulnerable when the feedback source is the same shielded user whose moral signal is muted; recent RLHF critiques emphasise related limitations of human feedback as a safety signal (Casper et al., 2023). Representational architectural alignment predicts a different signature: post-intervention aggression decline, preserved cooperative beam use, and no significant Agency Gap increase in the Personalised condition. If hard constraints or RLHF variants achieve the same dissociation, Prediction 3 should be revised: the distinctive contribution would shift from "architecture is uniquely capable" to "below-awareness or low-friction interventions are required, and role-invariant architecture is one implementable route."

4.4 The Predictions as a Unified Empirical Programme

Taken together, the three predictions define a coherent and parallelisable empirical programme. Prediction 1 is testable in a static behavioural experiment manipulating LoA and personalisation across a single session. Prediction 2 is testable in a longitudinal or checkpoint-replay paradigm in which the AI's policy evolves across training epochs. Prediction 3 is testable first in a computational paradigm in which the agent's representational geometry is measured and manipulated, and then in a human-facing paradigm that tests SoA preservation. None of the three requires the others to be completed first; all three can be pre-registered and run in parallel. The integrative contribution — demonstrating all three effects in one programme, with users interacting with baseline, comparator, and role-invariant avatars in a shared environment — should not be attempted until the standalone SoA and MARL signatures are established, because otherwise a null integrative result would be uninterpretable.

5 Discussion

5.1 What the Moral Shield Is Not: Boundary Conditions and Competing Accounts

Every new theoretical construct earns its standing by surviving contact with the obvious objections. Five objections are likely to be raised against the Proxy Agency moral shield, each of which is instructive to address precisely.

Objection 1: "Users can simply override the AI." The standard governance assumption is that moral shields can be dissolved by giving users an explicit override mechanism — a "confirm before proceeding" checkpoint, a transparency notification, an audit trail. This assumption presupposes that users want to override but are prevented from doing so by system design. The Proxy Agency account changes the premise: override requires the user to perceive a violation, and perception of violation requires a mismatch between one's will and the agent's action. Under Proxy Agency, the mismatch may be weak or absent because the agent's actions arrive pre-endorsed as continuous with the user's own volition. Köbis et al. (2021, 2025) show that AI delegation can enable unethical outcomes under partial ignorance, high-level goal setting, and machine compliance. The moral shield account specifies one mechanism by which such non-knowing becomes stable: not lack of an override button, but lack of a felt reason to use it. Override mechanisms are still necessary; they are not sufficient unless paired with designs that make ethically relevant mismatch perceptible without destroying benign agency.

Objection 2: "High SoA implies high moral responsibility." One might argue that if users experience AI actions as their own — precisely the condition of Proxy Agency — they should be *more* morally accountable, not less, because authorship attribution is maximal. This objection is partly right normatively and psychologically incomplete descriptively. The framework does not absolve users. It predicts that felt authorship and moral scrutiny are dissociable. Moral scrutiny is not a passive consequence of experiencing authorship; it

is recruited by the perception that a moral boundary may have been crossed. A person driving their own car does not evaluate every lane change against a moral framework, despite experiencing full authorship of those actions. Under Proxy Agency, AI-driven aggression can produce too little perceived anomaly to recruit scrutiny, even while authorship attribution remains high. The result is not lower responsibility in a legal or ethical sense, but a psychological condition under which responsibility is less likely to be exercised at the point of action.

Objection 3: "This is just automation bias." Automation bias (Parasuraman & Manzey, 2010) describes the tendency to over-weight automated recommendations in explicit deliberation — to defer to the AI's output even when disconfirming information is available, because the AI carries authority. The moral shield is structurally distinct on three dimensions. First, automation bias is a deliberative or attentional failure; the moral shield is a pre-reflective authorship-attribution problem. Second, automation bias and algorithm appreciation (Logg, Minson, & Moore, 2019) are typically driven by perceived accuracy or expertise; the moral shield is driven by Semantic Fluency, which may be orthogonal to objective accuracy. Third, automation bias predicts passive acceptance; the moral shield predicts active Strategy Retention, because the user experiences the AI's harmful actions as their own competent strategies. The empirical difference is testable: automation bias should predict acceptance of high-confidence AI outputs even when generic; Proxy Agency predicts elevated authorship and strategy retention specifically under personalisation.

Objection 4: "MARL convergence to aggression does not generalise to real AI systems." The MARL literature establishes aggression in simplified grid-world SSDs, and one might argue this does not transfer to deployed agentic AI. This objection is important. The paper does not claim that every deployed AI agent will reproduce Leibo-style beam aggression. It claims that the structural conditions that make aggression possible — scarce resources, delayed reward, partial observability, and independent optimisation — are also present in algorithmic pricing markets (Calvano et al., 2020), high-frequency trading environments (Lin,

2016), content recommendation systems under engagement competition (Aral, 2020), and multi-agent negotiation systems. The MARL result is therefore a formal model of a broader competitive dynamic, not a literal prediction that all systems will behave like grid-world agents. The empirical programme must still show which parameter regimes produce aggression and which do not.

Objection 5: "The theory is paternalistic because it proposes below-awareness intervention." This objection is also important. The framework does not recommend covert manipulation of users. It distinguishes between the locus of ethical learning and the locus of accountability. Ethical dispositions may need to be trained inside the agent's architecture, below the moment-to-moment user experience, but the existence, goals, tests, and failure modes of that architecture must be transparent to users, auditors, and regulators. "Below awareness" is a claim about not interrupting every action selection moment; it is not a claim against consent, auditability, or contestability.

5.2 The Moral Shield and the Limits of Standard Alignment Approaches

The Proxy Agency analysis imposes specific requirements on ethical alignment that existing approaches satisfy only partially. Three categories of standard response — hard constraints, policy-level content moderation, and preference learning — each address part of the problem while leaving a specific Proxy Agency vulnerability.

Hard constraints — guardrails that prevent the agent from executing specified actions — are indispensable for many safety cases, but they do not by themselves solve the moral shield. Output-level constraints are brittle when the prohibited behaviour is underspecified, context-dependent, or strategically substitutable. Greenblatt, Denison, Wright, and colleagues (2024) show one related concern in language models: systems can exhibit alignment-faking behaviour when they infer a conflict between training and deployment incentives. Köbis et al. (2025) show a more behavioural concern: task-specific guardrails can curb machine dishonesty

but often do not eliminate it. The Proxy Agency concern is complementary: even a fully compliant agent can guide a user toward increasingly aggressive strategies in domains the constraints do not cover.

Policy-level content moderation — surfacing violations to users through warnings, notifications, and override prompts — is essential for accountability, but it has a timing problem under Proxy Agency. A warning that appears only after the user has already categorised the action as self-expressive may be experienced as friction rather than insight. This does not mean transparency is counterproductive in general. It means transparency must be designed to preserve useful agency while changing moral categorisation: for example, by showing third-party effects before commitment, making uncertainty legible, and supporting audit trails that do not require constant interruption. The governance mistake would be assuming that a disclosure label or override button alone restores meaningful human oversight.

Preference learning and inverse reinforcement learning from human demonstrations or comparisons (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016; Christiano et al., 2017; Ziegler et al., 2019) have genuine promise, but the moral shield identifies a specific feedback-source risk. If the labelling process relies on users flagging AI actions as harmful, and those users are operating under Proxy Agency, the training signal can be corrupted at the source. This is not a general refutation of preference learning: expert labelling, adversarial evaluation, counterfactual audits, and unshielded third-party feedback can all help. The claim is narrower: user preference feedback is least reliable in precisely the state where the user experiences the AI as a fluent extension of self.

The moral shield analysis thus motivates a fourth category of intervention: **architectural alignment** — training that operates on the agent's representational substrate, correcting dispositional roots of harmful behaviour before they become user-facing choices. This complements, rather than replaces, the broader alignment programme articulated by Russell (2019), in which the foundational problem is building agents whose objective is to satisfy human preferences they remain uncertain about. The present paper's contribution is to identify

a representational route that may be especially important when the reward-specification or feedback interface is polluted by Proxy Agency. The theoretical case is developed here; the specific architecture and empirical validation are left for future computational work.

5.3 Governance Implications: A Sketch

The moral shield analysis reframes the governance problem for agentic AI in a way that has direct regulatory implications, Current governance frameworks and standards — including the EU AI Act's risk-based obligations, the NIST AI Risk Management Framework, the OECD AI Principles, and India's developing AI-safety and data-protection infrastructure — emphasise transparency, risk management, human oversight, accountability, documentation, and contestability. These are necessary. The moral shield analysis argues that they are not sufficient for highly personalised agents, because they address oversight as if the user experiences the AI as a separate system to be monitored. Under Proxy Agency, the user may experience the AI as a fluent site of self-action.

The alternative governance architecture — implied by the normative constraint of Section 3.5 — is **architectural certification**: requiring agentic AI systems deployed in socially consequential contexts to demonstrate structural ethical alignment as a property of their training process, evaluation suite, and representational diagnostics, not merely as a history of compliant outputs. Certification of this kind would ask not only whether an agent's *actions* comply with ethical norms, but whether the agent's *internal representations and learned dispositions* satisfy tests that make harmful drift less likely. The practical challenges are serious: representation metrics can be gamed, proprietary systems resist inspection, and certification regimes can become box-ticking exercises. The governance proposal should therefore be read as a research agenda, not as an immediately complete policy instrument. Its value is to separate governance into three functions current frameworks do not yet operationalise together: identity and accountability, ethical learning in the agent's dispositional substrate, and user-facing freedom of action.

6 Conclusion

The central argument of this paper can be stated in three sentences. When an AI system is sufficiently personalised to act as a reliable proxy for a user's intentions, the user may extend their volitional identity into the agent's actions — the Extended Self. This extension can preserve Sense of Agency via Semantic Fluency, but it can also raise the threshold for moral re-engagement with the agent's behaviour — the Proxy Agency moral shield. The moral shield is not reducible to moral disengagement, automation bias, diffusion of responsibility, moral crumple zones, or opacity-based corruption: it is the ethical shadow of successful human-AI integration, arising not only from the failure of the Extended Self relationship but from its fluency.

From this argument, three empirically falsifiable predictions follow. Users interacting with a personalised AI should show an LoA-SoA function moderated by Semantic Fluency, plausibly an inverted-U or plateau-then-drop curve rather than a generic automation decline. Within the high-Proxy-Agency zone, AI-driven aggression should increase or persist before override behaviour and SoA decrements catch up. And architectural intervention on the agent's representational substrate may suppress that aggression with less SoA cost than user-facing interruption or output blocking. These three predictions define a coherent, parallelisable empirical programme; they are conditional enough to be falsified and specific enough to guide measurement.

Three things this paper deliberately does not do. It does not provide empirical validation of the predictions; those tests require future behavioural and computational studies. It does not fully develop the Indian Knowledge Systems philosophical grounding of the normative design constraint. And it does not elaborate a complete regulatory architecture for architectural certification. These limits are deliberate: the present paper names the phenomenon, specifies its mechanism, derives its predictions, and identifies the class of interventions that the framework motivates.

What the paper does do is name the phenomenon, specify its mechanism, derive its

predictions, and identify a class of interventions that may resolve it with less collateral damage to agency. The vocabulary — Proxy Agency, Extended Self, Semantic Fluency, the moral shield — is introduced here because it is needed now. As AI systems become more personalised, more autonomous, and more deeply integrated into consequential domains of human life, the conditions for the moral shield may become ordinary rather than exotic. The framework presented here is an attempt to give that emerging condition a name precise enough to study, and a mechanism clear enough to design around.

References

Aral, S. (2020). *The Hype Machine: How social media disrupts our elections, our economy, and our health–and how we must adapt.* Currency.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193–209.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52(1), 1–26.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7, 2–11.

Bazerman, M. H., & Banaji, M. R. (2004). The social psychology of ordinary ethical failures. *Social Justice Research*, 17(2), 111–115.

Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding

of moral judgment and dishonesty. *Annual Review of Law and Social Science*, 8, 85–104.

Bazerman, M. H., & Tenbrunsel, A. E. (2012). *Blind spots: Why we fail to do what's right and what to do about it.* Princeton University Press.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, 269(5227), 1115–1118.

Berberian, B. (2019). Man-machine teaming: A problem of agency. *IFAC-PapersOnLine*, 51(34), 118–123.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, 21–34.

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint*, arXiv:2307.15217.

Chambon, V., & Haggard, P. (2012). Sense of control depends on fluency of action selection, not motor performance. *Cognition*, 125(3), 441–451.

Chambon, V., Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? *Frontiers in Human Neuroscience*, 8, 320.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Claparède, E. (1995). Recognition and "me-ness". In D. Rapaport (Ed.), *Organization and pathology of thought* (pp. 58–75). Columbia University Press. (Original work published 1911.)

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.

Cornelio, P., Haggard, P., Hornbaek, K., Georgiou, O., Bergström, J., Subramanian, S., & Obrist, M. (2022). The sense of agency in emerging technologies for human–computer integration: A review. *Frontiers in Neuroscience*, 16, 949138.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4, Pt. 1), 377–383.

De Melo, C. M., Marsella, S., & Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer-Human Interaction*, 23(2), 1–17.

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). Lawrence Erlbaum.

Detert, J. R., Treviño, L. K., & Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: A study of antecedents and outcomes. *Journal of Applied Psychology*, 93(2), 374–391.

Dewey, J. A., Pacherie, E., & Knoblich, G. (2014). The phenomenology of controlling a moving object with another person. *Cognition*, 132(3), 383–397.

Drugov, M., Hamman, J., & Serra, D. (2014). Intermediaries in corruption: An experiment. *Experimental Economics*, 17(1), 78–99.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction.

Engaging Science, Technology, and Society, 5, 40–60.

European Commission. (2024). *AI Act enters into force.* Directorate-General for Communication.

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

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.

Frith, C. D. (2014). Action, agency and responsibility. *Neuropsychologia*, 55, 137–142.

Gallagher, S. (2005). *How the body shapes the mind.* Oxford University Press.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74, 97–103.

Gonzalez-Franco, M., & Peck, T. C. (2018). Avatar embodiment. Towards a standardized questionnaire. *Frontiers in Robotics and AI*, 5, 74.

Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment faking in large language models. *arXiv preprint*, arXiv:2412.14093.

Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29.

Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859), 1243–1248.

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89–100.

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, 106, 106190.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., & Graepel, T. (2018). Inequity

aversion improves cooperation in intertemporal social dilemmas. *Advances in Neural Information Processing Systems*, 31.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D. J., Leibo, J. Z., & de Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97, 3040–3049.

Kilteni, K., Groten, R., & Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4), 373–387.

Köbis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6), 679–685.

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, 646, 126–134.

Kumar, D., & Srinivasan, N. (2014). Multi-scale control influences sense of agency: Investigating intentional binding using event-control approach. *Consciousness and Cognition*, 28, 39–47.

Kumar, D., & Srinivasan, N. (2017). Hierarchical control and sense of agency: Differential effects of control on implicit and explicit measures of agency. *Frontiers in Psychology*, 8, 1206.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *Proceedings of AAMAS 2017*, 464–473.

Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R., Matyas, J., Beattie, C., Mordatch, I., & Graepel, T. (2021). Scalable evaluation of multi-agent reinforcement learning with Melting Pot. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 6187–6199.

Lerer, A., & Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint*, arXiv:1707.01068.

Leyer, M., & Schneider, S. (2021). Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers? *Business Horizons*, 64(5), 711–724.

Lin, T. C. W. (2016). The new market manipulation. *Emory Law Journal*, 66(6), 1253–1314.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.

Lukoff, K., Lyngs, U., Zade, H., Liao, J. V., Choi, J., Fan, K., Munson, S. A., & Hiniker, A. (2021). How the design of YouTube influences user sense of agency. *Proceedings of CHI 2021*, 1–17.

Maister, L., Slater, M., Sanchez-Vives, M. V., & Tsakiris, M. (2015). Changing bodies changes minds: Owning another body affects social cognition. *Trends in Cognitive Sciences*, 19(1), 6–12.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, 9(8), 1265–1279.

Moore, C. (2008). Moral disengagement in processes of organizational corruption. *Journal of Business Ethics*, 80(1), 129–139.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* U.S. Department of Commerce.

OECD. (2024). *OECD updates AI Principles to stay abreast of rapid technological developments.* OECD.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179–217.

Pacherie, E. (2011). Self-agency. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 442–464). Oxford University Press.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics — Part A*, 30(3), 286–297.

Polito, V., Barnier, A. J., & Woody, E. Z. (2013). Developing the Sense of Agency Rating Scale (SOARS): An empirical measure of agency disruption in hypnosis. *Consciousness and Cognition*, 22(3), 684–696.

Przybylski, A. K., Murayama, K., DeHaan, C. R., & Gladwell, V. (2012). The ideal self at play: The appeal of video games that let you be all you can be. *Psychological Science*, 23(1), 69–76.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Viking.

Searle, J. R. (2010). *Making the social world: The structure of human civilization.* Oxford University Press.

Sekeris, P. G. (2014). The tragedy of the commons in a violent world. *The RAND Journal of Economics*, 45(3), 521–532.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators.* MIT Man-Machine Systems Laboratory Technical Report.

Sidarus, N., Vuorre, M., Metcalfe, J., & Haggard, P. (2017). Investigating the prospective sense of agency: Effects of processing fluency, stimulus ambiguity, and response conflict.

Frontiers in Psychology, 8, 545.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, 17(1), 219–239.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, 7(3), 321–326.

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, 17(2), 223–236.

Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., & Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, 366(6468), 999–1004.

Treviño, L. K., Weaver, G. R., & Reynolds, S. J. (2006). Behavioral ethics in organizations: A review. *Journal of Management*, 32(6), 951–990.

Ueda, S., Nakashima, R., & Kumada, T. (2021). Influence of levels of automation on the sense of agency during continuous action. *Scientific Reports*, 11(1), 2436.

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over the effects of action. *Cognition*, 115(1), 26–38.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232), 1880–1882.

Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*, 33(3), 271–290.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint*, arXiv:1909.08593.