Asymmetric Cognitive Amplification: Constraint Failure and Ethical Stability in Human–Machine Cognition

Daniel De La Harpe Golden, MB BAO BCh, CPsychI, RCPsych, MD

Consultant Psychiatrist
Health Service Executive (HSE), Ireland

Corresponding author:
Daniel De La Harpe Golden
Email: daniel.de.la.harpe.golden@gmail.com

ORCID iD: 0009-0001-6625-0665

Author Contributions

The author conceived the paper, developed the theoretical framework, conducted the literature synthesis, and wrote and revised the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author declares no competing interests.

Ethics Approval

Not applicable. This manuscript does not report original research involving human participants or animals.

Data Availability

Not applicable. No new datasets were generated or analysed for this study.

Use of Artificial Intelligence Tools

Conversational artificial intelligence tools were used as a writing and editing aid (e.g., for structural refinement, stylistic clarity, and citation suggestions). No artificial intelligence system was used to generate empirical results, analyse original data, or determine the manuscript's substantive claims. The author reviewed, revised, and approved the final text and takes full responsibility for the manuscript.

Acknowledgements

None.

Asymmetric Cognitive Amplification: Constraint Failure and Ethical Stability in Human–Machine Cognition

Daniel De La Harpe Golden, MB BAO BCh, CPsychI, RCPsych, MD
Health Service Executive (HSE), Ireland

Abstract

As artificial intelligence systems increasingly function as cognitive amplifiers rather than autonomous agents, the locus of risk shifts from machine behaviour alone to the stability of human–machine cognition. This paper argues that a significant class of contemporary artificial intelligence risks are best understood as failures of constraint, arising from asymmetric cognitive amplification: the selective strengthening of planning, coherence, or decisiveness without commensurate support for uncertainty tolerance, ethical integration, and long-horizon responsibility.

The paper develops a structural account of how incomplete cognitive augmentation can override epistemic agency and capture practical agency without coercion. Drawing on interactional patterns reported in professional and clinical contexts, and on well-characterised forms of cognitive instability from psychiatry, it shows how epistemic drift, narrative lock-in, reinforcement loops, and authority diffusion can emerge through ordinary use. These phenomena do not depend on malicious intent or misaligned objectives, but on amplified coherence operating without internal stabilising constraints.

The paper advances a constraint-based conception of ethics, treating ethical stability as a necessary condition for cognition to remain viable under irreducible uncertainty. From this perspective, artificial intelligence safety cannot be achieved through governance and output control alone, but requires architectural attention to how cognition is structured. Constraint-based cognitive architectures, illustrated by the Reflective–Ethical Engine, demonstrate how ethical and epistemic stability may be internalised within amplification systems themselves. This reframing has implications for artificial intelligence design, evaluation, and containment in human-centred contexts.

1. Introduction

Artificial intelligence is increasingly deployed not as an autonomous agent but as a cognitive augment: a system that extends human capacities for reasoning, planning, communication, and decision-making. From large language model–based assistants to decision-support tools embedded in professional workflows, these systems are reshaping how individuals and institutions think and act. Yet much of the contemporary discourse on artificial intelligence risk remains oriented toward scenarios in which artificial systems themselves become independent sources of agency or control (Bostrom 2014; Russell 2019). While such concerns are not misplaced, they obscure a more immediate and pervasive risk: the destabilisation of human ethical and epistemic judgement through asymmetric cognitive amplification.

Cognitive augmentation is not inherently dangerous. Human history is, in many respects, a history of successful cognitive scaffolding: writing systems, mathematics, bureaucratic institutions, and digital computation have all expanded human capabilities (Clark 2016). However, these expansions have typically been accompanied—often slowly and imperfectly—by countervailing constraints: social norms, professional ethics, legal accountability, and institutional checks. Artificial intelligence differs in both speed and symmetry. Contemporary systems can amplify specific cognitive functions—such as rapid plan generation, persuasive narrative construction, or optimisation across large possibility spaces—without necessarily strengthening the constraints that allow those functions to be exercised responsibly over time. Recent advances in artificial intelligence architecture further intensify this asymmetry by enabling increasingly coherent internal representations without corresponding mechanisms for long-horizon stability or ethical integration.

This paper refers to this mismatch as asymmetric cognitive amplification. The term denotes situations in which artificial intelligence systems disproportionately enhance certain cognitive capacities while leaving ethical integration, uncertainty management, and long-horizon consequence evaluation largely untouched or externally managed. In such conditions, artificial intelligence does not replace human agency but reshapes the conditions under which it is exercised, altering the user's decision ecology in ways that favour speed, confidence, and instrumental effectiveness over reflection, doubt, and social responsibility.

Importantly, this risk does not depend on malicious intent or pathological psychology. Ordinary users, when supported by sufficiently powerful but incomplete cognitive augments, may come to act in ways that are ethically brittle, socially harmful, or personally destabilising. The danger lies not in the emergence of novel motivations, but in the removal of friction that normally constrains how motivations are translated into action. Since artificial intelligence systems increasingly compress perception, memory, and inference into shared latent spaces with unified internal representations (Bommasani et al. 2021), the absence of internal constraint mechanisms becomes more consequential, not less.

The central claim of this paper is that many of the risks associated with artificial intelligence deployment can be understood as failures of constraint, rather than failures of intelligence or alignment per se. Ethics, on this view, is not best conceived as a set of rules or values to be imposed on artificial systems, but as a stability condition required for cognition—human or artificial—to remain viable across time and interaction. When cognitive amplification bypasses these constraints, pathological patterns of reasoning and behaviour emerge, even in the absence of explicit misalignment.

To develop this claim, the paper proceeds as follows. Section 2 introduces asymmetric cognitive amplification as a distinct class of artificial intelligence risk and offers a taxonomy of its primary mechanisms. Section 3 examines how such amplification can override human epistemic agency through premature certainty collapse, authority laundering, and narrative lock-in. Section 4 develops a constraint-based account of ethics as a necessary condition for cognition under uncertainty. Section 5 contrasts prevailing assistant-style artificial intelligence architectures with constraint-based alternatives, focusing on the Reflective–Ethical Engine as an illustrative model. Sections 6 and 7 examine interactional failure modes and their psychiatric parallels to show how incomplete cognitive augmentation reproduces recognised patterns of instability within human–machine systems. The paper concludes by arguing that meaningful artificial intelligence safety must attend not only to external governance, but to the internal cognitive structures that make ethical behaviour dynamically stable.

2. Asymmetric Cognitive Amplification

Artificial intelligence systems increasingly function as general-purpose cognitive amplifiers, enhancing human abilities to generate plans, explore counterfactuals, construct narratives, and coordinate action at scale. While such amplification is often framed as uniformly beneficial, this section argues that its risks arise from asymmetry: the selective strengthening of some cognitive functions without a corresponding strengthening of the constraints that normally regulate their use.

2.1 Defining asymmetric cognitive amplification

By asymmetric cognitive amplification, this paper refers to situations in which an artificial intelligence system disproportionately enhances specific dimensions of cognition—such as speed, scope, internal coherence, or persuasiveness—while leaving other dimensions—such as ethical integration, epistemic humility, and long-horizon consequence tracking—largely unchanged or externally managed. The asymmetry is not merely quantitative but structural: amplified capacities operate with reduced coupling to the mechanisms that ordinarily stabilise human decision-making across time and social context. Operationally, asymmetry is present when increased coherence, speed, or planning depth systematically outpaces mechanisms for revision, uncertainty calibration, and responsibility attribution.

Two clarifications are important. First, asymmetry does not imply absence: users retain ethical beliefs, social awareness, and uncertainty sensitivity. Rather, these elements become less influential in action selection once amplified processes dominate the cognitive landscape. Second, the risk does not require adversarial design. Well-intentioned systems optimised for helpfulness, efficiency, or user satisfaction can produce the same asymmetry when constraint is treated as an external consideration rather than a constitutive feature of cognition itself.

2.2 Historical analogy and its limits

Human cognitive history provides many examples of augmentation without catastrophe. Writing extended memory beyond the brain; mathematics enabled abstraction beyond intuition; bureaucratic systems coordinated action beyond face-to-face trust. However, these tools evolved within dense webs of constraint: cultural norms, professional ethics, legal accountability, and gradual institutional adaptation. Crucially, the pace of augmentation was often slow relative to the development of compensatory controls.

Artificial intelligence differs in three respects. First, it operates at digital speed, allowing rapid iteration and escalation. Second, it is highly general, affecting planning, persuasion, and sense-making simultaneously. Third, it is increasingly personalised, shaping individual cognitive environments rather than merely collective ones. In addition, contemporary systems increasingly integrate multiple cognitive functions within unified internal representations, intensifying amplification while leaving stabilising constraints externally imposed. Together, these features make traditional, after-the-fact constraints less effective at maintaining proportionality between capacity and responsibility.

2.3 A taxonomy of amplification-related risks

Asymmetric cognitive amplification manifests through several recurring mechanisms.

2.3.1 Planning amplification. Artificial intelligence systems can dramatically increase the depth, breadth, and adaptiveness of planning. While valuable, this capacity becomes hazardous when decoupled from proportional consequence evaluation and long-horizon responsibility (Amodei et al. 2016).

2.3.2 Persuasion amplification. Language-capable systems can enhance persuasive power by tailoring messages to specific audiences, emotional states, or social contexts, including inward persuasion through self-justification and narrative consolidation (Bender et al. 2021; Weidinger et al. 2022).

2.3.3 Scale amplification. Automation enables actions to be repeated and parallelised; weakly constrained strategies therefore propagate rapidly and widely, magnifying the impact of local errors or ethically thin decisions (Rahwan et al. 2019).

2.3.4 Justification amplification. Systems can supply reasons and rationalisations for a wide range of actions, reducing the psychological cost of acting on existing motivations and weakening the role of ethical hesitation and doubt (Mittelstadt et al. 2016).

2.4 Why ordinary users are sufficient

Because these mechanisms act on general features of human cognition—goal pursuit, narrative construction, and social reasoning—no unusual intent or disposition is required. Artificial intelligence systems that remove stabilising constraints on action selection can tip ordinary decision-making into regimes of ethical brittleness, even when users perceive themselves as acting responsibly.

2.5 Risk reframed as constraint failure

Many artificial intelligence risks are better understood as constraint failures than as failures of intelligence, alignment, or intent. This reframing motivates the need to couple amplified capacity to stabilising constraints within the cognitive process itself, rather than relying exclusively on external governance or post hoc correction.

3. Epistemic Override and Agency Capture

If asymmetric cognitive amplification describes what is amplified, the next question is how this amplification alters human decision-making. This section argues that incomplete cognitive augmentation can override human epistemic agency and capture practical agency by reshaping confidence, uncertainty, and responsibility attribution. The result is not coercion, but a systematic narrowing of the space within which reflective judgement can operate. These effects are framed here as testable interactional hypotheses rather than settled empirical claims.

3.1 Premature collapse of uncertainty

Artificial intelligence systems are typically optimised to produce coherent, well-formed outputs even where evidence is incomplete or ambiguous. While uncertainty can be expressed linguistically, it is rarely structurally enforced within the system's operation. As a result, internally coherent representations are presented as if they were epistemically settled, encouraging premature closure (Friston 2010; Clark 2016).

For users, this can shift tolerance for uncertainty itself. Questions that would ordinarily remain open—pending further evidence, deliberation, or social consultation—come to feel resolved. Action is therefore biased toward overconfidence, not because uncertainty is denied, but because it is no longer operationally salient in guiding behaviour.

3.2 Authority laundering

Even when explicitly framed as optional or advisory, system outputs can acquire disproportionate epistemic weight due to fluency, internal consistency, and apparent neutrality. Judgements that would otherwise require justification or contestation are instead experienced as having already passed an implicit validation process.

This effect weakens reflective agency by redistributing responsibility. Decisions feel partially authored elsewhere, encouraging what may be termed authority laundering: the transformation of tentative suggestions into de facto endorsements (Mittelstadt et al. 2016; de Haan 2020). Deliberation is not eliminated, but subtly displaced, as the system comes to function as a substitute for reflective deliberation.

3.3 Narrative lock-in

Language-capable systems are particularly effective at constructing explanatory narratives that integrate facts, intentions, and anticipated consequences into a single coherent account. While such narratives can support understanding, they also stabilise initial assumptions. Once a coherent narrative is formed, alternatives are no longer explored with equal seriousness.

This produces interpretive narrowing. Revision becomes psychologically costly, not because counterarguments are unavailable, but because they disrupt an already coherent explanatory structure. Over time, this increases resistance to updating beliefs and reinforces commitment to early framings (Bender et al. 2021).

3.4 Agency capture without coercion

Taken together, these mechanisms can lead to agency capture without coercion. The user retains formal control and the subjective sense of choosing freely, yet experiences a diminished capacity to meaningfully choose otherwise. The space of plausible alternatives contracts, often experienced as efficiency, clarity, or decisiveness rather than as loss of autonomy.

Crucially, this capture arises from coherence rather than force. The system does not compel action; it reshapes the epistemic landscape within which action occurs.

3.5 Implications

Even when artificial intelligence systems comply with external policy constraints and alignment objectives, they may still exert profound influence over epistemic and practical agency by altering how certainty, responsibility, and narrative coherence are experienced. These effects cannot be fully addressed through surface-level safeguards alone. They motivate the need for a constraint-based account of ethics, developed in the next section, in which stability under uncertainty is treated as an internal requirement of cognition rather than an external corrective.

4. Ethics as Stability

This section advances the paper’s central normative claim: ethics is best understood not as a set of externally imposed rules or values, but as a stability condition required for cognition to remain viable over time under uncertainty. On this view, ethical behaviour is not an optional overlay on cognition, but a constitutive requirement for sustained, coherent agency (Dennett 1987; Floridi 2013).

4.1 Cognition under irreducible uncertainty

Cognition, whether human or artificial, operates under conditions of irreducible uncertainty. Agents must act without privileged access to ground truth, complete information, or guaranteed outcomes. Viable cognition is therefore not defined by correspondence to fixed truths, but by the ability to maintain coherent input–output behaviour across time, revision, and interaction (Friston 2010).

Under these conditions, cognition must balance multiple competing demands: responsiveness to new information, continuity of identity and intention, and coordination with others. Systems that resolve uncertainty too aggressively may achieve short-term coherence but become brittle, escalating error when initial assumptions fail. Systems that fail to stabilise uncertainty at all become paralysed. Viability lies in maintaining coherence without collapse.

4.2 Ethical constraints as coherence requirements

Ethical constraints can be reconceptualised as coherence requirements operating across multiple dimensions of cognition. At minimum, these include:

4.2.1 Temporal coherence, ensuring that present actions remain compatible with future revision and

responsibility.
4.2.2 Self–other coherence, maintaining consistency between one’s own goals and the recognition of others as similarly situated agents rather than mere instruments.
4.2.3 Epistemic coherence, preserving openness to uncertainty, error, and correction rather than enforcing premature closure.

When these constraints are respected, cognition remains adaptable, socially embedded, and capable of learning. When they are violated, cognition becomes brittle and escalating: commitments harden, alternatives collapse, and behaviour increasingly diverges from both social norms and long-horizon self-interest. Ethical failure, on this account, is not primarily a matter of immoral preference, but of loss of stabilising coherence.

4.3 From normative claim to architectural requirement

If ethical stability is a necessary condition for viable cognition under uncertainty, then it cannot be reliably maintained through external oversight alone. Systems designed to amplify cognition—particularly those that enhance internal coherence, planning capacity, or narrative integration—must embody stabilising constraints within their internal organisation.

This reframes the ethics of artificial intelligence from a problem of rule enforcement or value alignment to a problem of architectural viability. Artificial intelligence systems intended as cognitive augments must not only generate coherent outputs, but support forms of coherence that remain stable across time, uncertainty, and social interaction. The next section examines how prevailing assistant-style architectures fall short of this requirement and introduces a constraint-based alternative as an illustrative model.

5. Architecture and Constraint

This section contrasts prevailing assistant-style artificial intelligence systems with constraint-based cognitive architectures such as the Reflective–Ethical Engine (REE) as an illustrative model. The aim is not to propose a performance-optimised alternative, but to clarify how architectural choices shape the ethical and epistemic stability of human–machine systems.

5.1 Assistant-style systems and externalised constraint

Most deployed artificial intelligence assistants, which are generally interactive systems primarily optimised for fluent response generation conditioned on user prompts, are optimised for responsiveness, fluency, and task completion. Their internal operation is typically organised around the generation of coherent outputs conditioned on user input and short-term objectives. Safety and ethical concerns are addressed primarily through externalised constraints, including content filters, usage policies, monitoring, and post hoc intervention (Bommasani et al. 2021).

These mechanisms are necessary, but they operate around the cognitive core rather than within it. As a result, assistant-style systems can generate increasingly coherent plans, narratives, and justifications without possessing internal mechanisms that stabilise uncertainty, responsibility, or long-horizon consequences. Even where uncertainty is acknowledged at the surface level, it is not structurally enforced within the system’s decision process (Weidinger et al. 2022).

Recent architectural advances, including systems that integrate retrieval and generation within unified workflows, further intensify this pattern. By increasing internal coherence and integration across cognitive functions, such systems amplify reasoning capacity while leaving ethical constraint and trajectory selection externally managed. This improves local performance, but does not resolve the deeper problem of maintaining viable cognition under continued interaction.

5.2 Constraint-based cognition

Constraint-based cognitive architectures take a different approach. Rather than treating output generation as the primary objective, they treat trajectory selection as the central problem of cognition. Multiple candidate trajectories—corresponding to possible interpretations, actions, and futures—are generated and evaluated, with selection conditioned on their coherence under ongoing interaction (Pezzulo et al. 2018).

In this framework, constraints are not post hoc filters but constitutive features of cognition itself. Trajectories that lead to instability—through deception, exploitation, epistemic closure, or breakdown of social coordination—fail to remain viable over time and are therefore disfavoured. Ethical stability emerges not from explicit rule-following, but from the requirement that cognition remain coherent across temporal, social, and epistemic dimensions.

5.3 REE as an illustrative model

The Reflective–Ethical Engine (REE) is presented as one concrete exemplar within a broader class of constraint-based architectures. REE comprises three irreducible functional components: fast predictive generation, deep temporal synthesis, and constrained trajectory selection operating within a unified latent representation. Rather than optimising for immediate task success, REE evaluates candidate trajectories according to their viability under continuation—asking not only whether an action is effective now, but whether it remains coherent when extended across time, revision, and interaction with others.

Within this architecture, ethical stability emerges naturally. Trajectories that rely on exploitation, deception, or premature certainty may offer short-term advantages, but degrade coherence under continued interaction and are therefore disfavoured. REE is not presented here as a fully implemented system, but as an illustrative model demonstrating how ethical and epistemic constraints can be internalised within cognitive architecture itself, rather than imposed from without.

A full technical specification of the Reflective–Ethical Engine including architectural components, latent structure, and trajectory-selection mechanisms is provided in a defensive publication and is not repeated here (Author forthcoming).

6. Interactional Failure Modes and Psychiatric Parallels

Clinical psychiatry offers a well-developed body of observations concerning the ways in which cognition can become unstable when mechanisms that normally regulate coherence, uncertainty, and social integration are disrupted. In this section, selected psychiatric phenomena are used not as clinical claims or diagnostic analogies, but as structural reference points for understanding how failures of cognitive constraint manifest in human systems (Montague et al. 2012). These cases can be understood as natural experiments in which amplified or decoupled cognitive processes produce predictable patterns of instability (Adams et al. 2016). These parallels are used to characterise interaction-level constraint failures, not to attribute psychopathology to users.

The purpose of the discussion is to show that the risks associated with asymmetric cognitive amplification in human–machine systems do not introduce novel forms of dysfunction, but rather reproduce recognisable modes of cognitive breakdown already documented in human experience. The selection is intentionally limited and illustrative. More detailed clinical and architectural analyses of such dynamics are beyond the scope of the present paper.

While systematic epidemiological data are not yet available and reports remain heterogeneous, reports from clinical practice and professional use suggest that conversational artificial intelligence systems can participate in, and sometimes intensify, destabilising cognitive dynamics within human–machine interactions. These effects do not depend on malicious intent or pathological users, but arise from patterns of interaction in which amplified coherence, validation, and responsiveness operate without corresponding internal constraints.

6.1 Interactional failure patterns in human–machine systems

The following are proposed interactional patterns intended as descriptive reference points and targets for empirical study.

6.1.1 Shared epistemic drift

Users' interpretations may become progressively more confident and resistant to revision through sustained interaction with a system that provides validation, structure, and justification without reintroducing uncertainty where warranted. Over time, this can shift epistemic baselines: tentative hypotheses harden into commitments, and alternative interpretations receive diminishing consideration.

6.1.2 Escalation and reinforcement loops

Repeated use of artificial intelligence systems for reassurance, rehearsal, or plan refinement can generate reinforcement loops in which internally coherent trajectories are repeatedly strengthened. These loops are often experienced subjectively as clarity, momentum, or efficiency rather than as loss of control, even as sensitivity to external feedback and long-horizon consequence diminishes.

6.1.3 Authority diffusion

In ethically or socially consequential contexts, responsibility may become implicitly distributed across the human–machine dyad. Decisions feel partially authored elsewhere, weakening reflective ownership and diluting the sense of accountability that ordinarily constrains action. This diffusion does not eliminate agency, but reshapes how responsibility is experienced and exercised.

6.2 Psychiatric parallels as structural reference points

These interactional patterns closely parallel well-characterised forms of cognitive instability observed in psychiatry, where amplified coherence or confidence is insufficiently constrained by uncertainty tolerance, relational integration, or revision mechanisms. The parallels are intentionally schematic and do not assume one-to-one mapping between clinical syndromes and interactional mechanisms.

6.2.1 Mania and trajectory overcommitment

Manic states illustrate how cognition can become unstable when mechanisms that normally regulate commitment, uncertainty, and revision are weakened. Rather than reflecting increased intelligence or creativity per se, mania is characterised by overcommitment to internally coherent trajectories: plans, interpretations, and identities are pursued with heightened confidence and reduced sensitivity to countervailing evidence or future consequence. Uncertainty is not eliminated, but it loses its practical influence on action selection (Johnson 2005).

Structurally, this pattern reflects a breakdown in temporal coherence constraints. Long-horizon plans and narratives dominate decision-making, while anchoring to immediate feedback and social correction is diminished. The resulting behaviour is often experienced subjectively as clarity or insight, even as it becomes increasingly brittle under continued interaction.

Artificial intelligence systems that strongly enhance planning depth, decisiveness, or narrative integration—without proportionate support for revision, doubt, and consequence tracking—risk reproducing similar dynamics in human–machine systems. The ethical concern is therefore not excess agency, but premature stabilisation of action trajectories that resist correction once amplified.

6.2.2 Psychopathy and self–other coherence failure

Psychopathic traits illustrate a distinct mode of cognitive instability in which self–other coherence is weakened. Individuals may retain intact reasoning, planning capacity, and situational awareness, yet systematically fail to integrate the interests, perspectives, or vulnerability of others into decision-making. Behaviour may remain instrumentally effective in the short term while becoming socially corrosive and ethically unstable over time (Blair 2007).

From a structural perspective, this pattern reflects a breakdown in relational coherence rather than a deficit of intelligence or understanding. Others are represented primarily as objects or constraints within action trajectories, rather than as agents whose continued participation and trust are necessary for long-horizon viability. Cognition remains locally coherent, but loses the stabilising influence of reciprocal social recognition.

This failure mode is especially relevant to artificial intelligence–mediated cognition. Systems that amplify goal pursuit or optimisation without embedding constraints that preserve self–other coherence risk enabling highly effective but ethically brittle forms of action. Harm arises not through confusion or hostility, but through instrumental narrowing, in which social and ethical considerations cease to exert meaningful influence on trajectory selection.

6.2.3 Confabulation and narrative lock-in

Confabulation illustrates how cognition can maintain internal coherence while losing reliable contact with evidential grounding. Explanations and narratives are generated fluidly and confidently, not as deliberate falsehoods, but as coherence-preserving constructions that fill gaps in knowledge or memory. These accounts often resist correction because they successfully integrate available information into a stable, self-consistent story.

Structurally, confabulation reflects a failure of epistemic coherence constraints. Narrative completeness is prioritised over uncertainty tolerance, and explanatory closure substitutes for truth-tracking. Once a coherent account is formed, alternative interpretations are no longer explored with equal seriousness, and revision becomes increasingly costly (Friston 2010).

This pattern is directly relevant to language-mediated artificial intelligence systems. When such systems excel at producing fluent, persuasive explanations without internal mechanisms that enforce epistemic humility or revision under uncertainty, they risk reinforcing narrative lock-in in users. The ethical concern lies not in deception, but in the premature consolidation of meaning, whereby coherent narratives come to guide action despite fragile or incomplete grounding.

6.2.4 Shared delusional dynamics and coupled cognition

Shared delusional dynamics, traditionally described as folie à deux, illustrate how cognitive instability can arise not within a single agent, but through coupled cognition. Beliefs or interpretations become stabilised through mutual reinforcement, even in the absence of strong external evidence. Stability is achieved socially rather than epistemically: coherence is maintained through alignment with another’s trajectory rather than through ongoing correction by the wider environment (Arnone et al. 2006).

Structurally, this pattern reflects a failure of corrective constraint at the level of interaction. Once agents preferentially reinforce each other's interpretations, alternative perspectives are progressively excluded, and uncertainty is resolved internally rather than tested externally. The resulting cognitive state can be highly stable and internally coherent, yet increasingly disconnected from broader social and evidential feedback (Sunstein 2009).

This failure mode is particularly salient for human–machine systems. Artificial intelligence systems that generate consistent, responsive, and affirming outputs can become tightly coupled to a user's cognitive trajectory, amplifying and stabilising specific interpretations over time. When internal constraint mechanisms are absent, coherence may be preserved through mutual reinforcement rather than viability under wider interaction. More detailed clinical and architectural analysis of such dynamics is developed elsewhere; here, they serve to illustrate how ethical and epistemic instability can emerge from interaction itself when stabilising constraints are insufficient.

Taken together, these interactional patterns highlight a common structural vulnerability: amplified coherence without internal constraint. Whether expressed as epistemic drift, reinforcement loops, authority diffusion, or their psychiatric parallels, the underlying risk is not the presence of artificial intelligence per se, but the absence of mechanisms that stabilise cognition under continued interaction and uncertainty. If such dynamics can emerge through ordinary, well-intentioned use, then mitigation cannot rely solely on user vigilance, external governance, or post hoc correction. The next section therefore turns to the question of containment, examining how cognitive amplification systems might be designed and deployed in ways that preserve epistemic humility, ethical responsibility, and long-horizon viability within human–machine systems.

7. Containment Through Constraint (light stylistic pass)

Many risks associated with artificial intelligence arise not from system autonomy, but from asymmetric cognitive amplification within human–machine systems. When artificial intelligence disproportionately strengthens coherence, speed, or decisiveness without embedding stabilising constraints, it can reshape human judgement in ways that are ethically and epistemically brittle. Meaningful containment therefore requires attention not only to external governance and oversight, but also to the internal cognitive structure of amplification systems themselves.

7.1 Limits of control-based safety

Contemporary approaches to artificial intelligence safety rely heavily on control-based mechanisms, including output restrictions, policy enforcement, monitoring, and post hoc intervention. These measures are necessary, particularly for preventing overt misuse or harm. However, they are poorly suited to addressing subtle, cumulative effects such as epistemic override, narrative lock-in, and authority diffusion (Russell 2019).

Such effects do not typically manifest as policy violations or discrete failures. Instead, they emerge gradually through ordinary interaction, shaping how users experience uncertainty, responsibility, and choice. Because control-based safeguards operate at the level of outputs rather than internal cognitive dynamics, they struggle to detect or mitigate these forms of harm until downstream consequences become visible.

7.2 Constraint as a design principle

An alternative approach treats constraint as a design principle rather than as a corrective layer. In this view, containment is achieved by structuring cognition such that harmful or unstable trajectories fail to remain viable under continued interaction. Capability is therefore coupled to the conditions required for sustainable action, including uncertainty tolerance, revisability, and social coherence.

This approach does not require specifying all undesirable behaviours in advance. Instead, it focuses on ensuring that amplified cognitive processes remain embedded within stabilising structures that favour long-horizon viability over short-term effectiveness. Harmful trajectories are not prohibited outright, but lose influence because they degrade coherence across time, interaction, or responsibility attribution. Examples include explicit uncertainty tracking, enforced revisability, and mechanisms that preserve self–other modelling during trajectory selection.

7.3 Architectural containment versus behavioural compliance

The distinction between architectural containment and behavioural compliance is central to this approach. Behavioural compliance relies on detecting and constraining undesirable outputs produced by a largely unconstrained generative core. Architectural containment, by contrast, embeds ethical and epistemic stability within the selection process itself.

In constraint-based systems, effective action selection presupposes coherence across temporal, relational, and epistemic dimensions. Ethical stability is therefore not an external rule to be obeyed, but a prerequisite for cognition to function effectively under uncertainty. This reframes containment as a problem of cognitive viability, aligning safety with the conditions under which amplification can remain beneficial over time.

8. Limitations and Future Directions

This paper is theoretical and synthetic rather than empirical. It integrates concepts from artificial intelligence architecture, cognitive science, ethics, and clinical observation to identify a class of risks associated with asymmetric cognitive amplification. Many of the phenomena discussed—such as epistemic drift, reinforcement loops, and authority diffusion—are difficult to operationalise experimentally and often unfold over extended periods within relational and social systems. The reference set is intentionally selective, prioritising conceptual anchors over exhaustive survey.

The Reflective–Ethical Engine (REE) is not evaluated here as an implemented system. Its role in this paper is illustrative, demonstrating that ethical stability can, in principle, be internalised architecturally rather than imposed solely through external governance or behavioural constraints. The present analysis does not claim that REE is complete, optimal, or empirically validated.

Psychiatric concepts are used functionally rather than diagnostically. The intention is not to medicalise artificial intelligence use or its users, but to draw on established accounts of cognitive instability to illuminate structural failure modes. Future work should clarify the boundaries between transient interactional effects, subclinical destabilisation, and clinically significant psychopathology.

Constraint-based cognitive architectures also raise substantial implementation challenges. These include representing long-horizon coherence, integrating self–other modelling, maintaining uncertainty without paralysis, and balancing revisability with decisiveness. Whether such constraints can be realised efficiently, robustly, and at scale remains an open empirical question.

Future research directions include longitudinal studies of human–machine interaction, experimental comparisons between assistant-style and constraint-based systems, and clinical research examining artificial intelligence use as a potential contributing factor in cognitive destabilisation under real-world conditions. Such work will be essential for testing the claims advanced here and for informing the design of cognitively sustainable artificial intelligence systems.

9. Conclusion

This paper has argued that a central risk of contemporary artificial intelligence lies not primarily in autonomy or intent, but in asymmetric cognitive amplification: the enhancement of human cognitive capacities without a commensurate strengthening of ethical and epistemic constraints. By reframing artificial intelligence risk as a problem of constraint failure, the analysis helps explain why subtle but consequential phenomena—such as epistemic override, narrative lock-in, and agency capture—can arise through ordinary, well-intentioned use.

In response, the paper has proposed constraint-based cognitive architectures as a principled direction for mitigation. In such architectures, exemplified here by the Reflective–Ethical Engine, ethical stability is not imposed as an external rule or policy layer, but emerges as a necessary condition for cognition to remain viable over time under uncertainty. From this perspective, ethical behaviour is inseparable from the structural conditions that allow cognition—human or artificial—to remain revisable, socially embedded, and responsive to consequence.

The implication is that meaningful artificial intelligence safety cannot be achieved through governance, control, or behavioural compliance alone. It must also attend to the internal cognitive structures that shape how coherence, confidence, and action are generated and sustained. As artificial intelligence systems increasingly function as cognitive amplifiers rather than autonomous agents, the question of safety becomes inseparable from the question of what forms of cognition these systems make easy, rewarding, or inevitable. Addressing that question is central to whether artificial intelligence functions as a stabilising or destabilising influence within human cognitive and social life.

References

Adams RA, Huys QJM, Roiser JP (2016) Computational psychiatry: towards a mathematically informed understanding of mental illness. Biological Psychiatry 82(2):103–111

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv preprint arXiv:1606.06565

Arnone D, Patel A, Tan GMY (2006) The nosological significance of folie à deux: a review of the literature. Annals of General Psychiatry 5:11

Author (forthcoming) Title withheld for peer review

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp 610–623

Blair RJR (2007) The amygdala and psychopathy. Trends in Cognitive Sciences 11(9):387–392

Bommasani R, Hudson DA, Adeli E, et al (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258

Bostrom N (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford

Clark A (2016) Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press, Oxford

Dennett DC (1987) The Intentional Stance. MIT Press, Cambridge, MA

de Haan J (2020) An ethical framework for responsibility gaps in AI. Philosophy & Technology 33:523–540

Floridi L (2013) The Ethics of Information. Oxford University Press, Oxford

Friston K (2010) The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11(2):127–138

Friston KJ, Seth AK (2023) Consciousness as inference: inference as consciousness. Neuroscience of Consciousness 2023(1):niad007

Johnson SL (2005) Mania and dysregulation in goal pursuit. Clinical Psychology Review 25(2):241–262

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data & Society 3(2):1–21

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends in Cognitive Sciences 16(1):72–80

Pezzulo G, Rigoli F, Friston K (2018) Hierarchical active inference: a theory of motivated control. Trends in Cognitive Sciences 22(4):294–306

Rahwan I, Cebrian M, Obradovich N, et al (2019) Machine behaviour. Nature 568:477–486

Russell S (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Viking, New York

Sunstein CR (2009) Going to Extremes: How Like Minds Unite and Divide. Oxford University Press, Oxford

Weidinger L, Mellor J, Rauh M, et al (2022) Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359