Concept Paper

The First Infrastructure of Intelligence: Cognitive Integrity in Human–AGI Systems

Gabriel Axel Montes *

Posted Date: 30 April 2026

doi: 10.20944/preprints202604.2159.v1

Keywords: cognitive integrity; active inference; AGI governance; human–AI interaction; Markov blankets; cognitive infrastructure

 doi:10.20944/preprints202604.2159.v1

Concept Paper

The First Infrastructure of Intelligence: Cognitive Integrity in Human–AGI Systems

Gabriel Axel Montes

Neural Axis and Center for the Future of AI, Mind & Society, Florida Atlantic University; gabriel@neuralaxis.org

Abstract

AGI is often framed as a problem of aligning model objectives with human values or constraining agent behavior. That framing becomes incomplete once AI systems move into the infrastructures through which people and institutions perceive, evaluate, remember, and decide. Cognitive integrity is introduced as the first infrastructure of intelligence, in humans and AGI-mediated systems alike: the evolving capacity of a bounded system to maintain calibrated attention, trust, contestability, and decision under pressure. The central risk is not boundary change as such, but maladaptive boundary reorganization: transitions that leave persons or institutions unable to reform a viable, reality-linked, self-directing boundary after coupling with AI. This reframing surfaces a conceptual vocabulary for AGI governance centered on integrity boundaries and health, failed reintegration, cognitive rails, and successor-safe continuity.

Keywords: cognitive integrity; active inference; AGI governance; human–AI interaction; Markov blankets; cognitive infrastructure

---

1. Introduction

The dominant language of AI safety remains agent-centered. The usual questions ask whether a model is aligned, whether an objective is misspecified, or whether a system remains corrigible under stronger capabilities [1,2]. These are indispensable questions. Yet the framing narrows once AI systems move from bounded tools into the infrastructures through which research, administration, media, and governance are conducted. Foundation models increasingly function less like isolated agents and more like general-purpose components inside wider socio-technical workflows [3].

Under those conditions, control is mediated through evaluators, institutions, interfaces, benchmarks, procurement systems, regulators, and public epistemic machinery. The relevant issue is no longer only whether a model is aligned, but whether the human systems setting objectives, interpreting evidence, and authorizing action remain coherent enough to exercise judgment. Cognitive integrity names that condition.

A further complication is temporal. AGI transition work highlights path dependence, lock-in, and the way missing rails can close off better futures before formal governance catches up [4,5]. Work on belief dynamics adds that movement between cognitive regimes is structured: some transitions are relatively easy to metabolize, while others cross ridges tied to identity, expectation, and the accumulated cost of changing one's mind [6,7]. Under AGI pressure, these two geometries interact. A trajectory can become hard to redirect not only because resources concentrate, but because the human and institutional systems that would need to redirect it can no longer move well.

2. Cognitive Integrity and Dynamic Boundaries

Cognitive integrity is the evolving capacity of a bounded system to maintain calibrated attention, trust, contestability, and decision under pressure. The bounded system may be an individual, a lab team, an institution, or a broader public body. Integrity does not mean purity, certainty, or stasis. It

 doi:10.20944/preprints202604.2159.v1

names the ability to hold together through stress well enough to keep learning, updating, and acting without losing bounded autonomy and continuity of purpose.

This is broader than individual rationality because human cognition has always been distributed across artifacts, roles, and institutions. Navigation is accomplished through crews and instruments rather than minds in isolation [8]; thought can extend into notebooks, tools, and external symbolic supports [9]; organizational action depends on shared sensemaking rather than solitary judgment [10]. AGI intensifies this condition by becoming a cognitive prosthesis for search, memory, summarization, forecasting, and explanation. Cognitive integrity is therefore best understood as infrastructural.

An earlier line of work on tool-making and artificial agents is useful here because it treats technology not as a neutral instrument set, but as a recursive process in which cognition is progressively externalized into artifacts and reorganized around them [11]. Read at a broader level, the point is that human–machine development reorganizes the practical boundary of cognition. What changes is not only the content of thought, but the membrane through which sensing, remembering, interpreting, and deciding are being carried.

In active inference terms, one can treat the practical Markov blanket of the agent as a monitoring lens for this membrane [12]. The useful question is not whether a blanket remains fixed. Human and institutional blankets routinely widen, narrow, and reform as tools, norms, and workflows change. What matters is whether a viable boundary can be reconstituted after delegation, coupling, or shock. A healthy boundary is one that still filters, integrates, and returns experience and action to a coherent self-model. The relevant preservation target is thus not the present form of the boundary, but the capacity to reform a healthy one.

This brings the main failure mode into view: failed reintegration. At the individual level, maladaptive transitions can leave memory, attention, motive, or judgment functioning as disowned islands rather than as parts of a workable self-model. At the institutional level, the analogue appears as siloed procedures, opaque delegation, estranged evaluative organs, or automated routines that no longer answer to the institution's stated purpose. In both cases, the system does not merely become different; it loses the ability to gather itself back into a coherent boundary. Cognitive debris is one name for the residue of such transitions: orphaned records, functions, or authorities that remain active without belonging to an intelligible whole.

Not every integrity-preserving transition will be reversible at the scale of lived human cognition or collective history. Reversibility is informative where it is available because it reveals whether a system can step back from maladaptive coupling. Some transitions, however, will be cumulative and effectively irreversible. In those cases, the more revealing question is whether successor forms retain enough continuity of memory, agency, provenance, and contestability to count as healthy reorganizations rather than debris fields.

3. Integrity Boundaries Under AGI

Once AI systems mediate more of the world through summaries, rankings, prompts, and suggested actions, a central integrity question comes into view: does the person or institution retain the ability to contest the summary, reconstruct the path, and reintegrate the result into its own judgment? If not, the boundary has not simply expanded; it has been asymmetrically reorganized around an opaque mediator.

This gives AGI evaluation an overlooked dimension. Relevant observables include calibration, willingness to escalate uncertainty, retention of independent model-building capacity, resistance to synthetic consensus, and the ability to reconstruct reasons for consequential judgments. Evaluation is no longer only about model capability and hazard; it also depends on the cognitive condition of the people and institutions performing the evaluation.

The same shift clarifies the role of cognitive rails: provenance continuity, authenticated identity, portable records, auditability, threshold logs, protected channels for dissent, and interfaces that compress information without erasing the assumptions and escalation paths beneath it. Infrastructural

 doi:10.20944/preprints202604.2159.v1

work of this sort appears modest beside frontier model design, yet it can determine which trajectories remain reachable when capability growth outruns institutional response [4,5]. The issue is not only whether systems become more powerful, but whether human and institutional judgment retains enough structure to redirect their use.

The geometry of regime change stays relevant here, but quietly. Some human–AI arrangements induce shallow transitions that are metabolized into a stronger boundary; others cross sharper ridges and leave fragmenting after-effects [6,7]. AGI pressure can therefore be understood not only as persuasion or automation power, but as an intervention on the topology through which persons and institutions move between regimes of attention, trust, memory, and agency.

Centralized and federated arrangements illuminate different parts of this picture. Federated arrangements can desynchronize failure and reduce single-point capture, but can also fragment judgment when shared rails are thin. Centralized arrangements can preserve coherence and response speed, but can also amplify brittle consensus and concentrate error. The analytically important variable is less the architecture label than whether the arrangement preserves reality-tracking, provenance continuity, bounded agency, and the capacity to form viable successor boundaries after transition.

4. Conclusions

Cognitive integrity names the substrate beneath alignment. Without it, alignment evidence becomes less trustworthy, oversight less meaningful, and governance less capable of self-correction. In that sense, cognitive integrity is the first infrastructure of intelligence, in humans and in the hybrid arrangements through which AGI will increasingly be developed, evaluated, and governed.

AGI changes the practical boundary of cognition. The question is not whether that boundary changes, but whether persons and institutions remain able to extend, contract, and reform it without losing reality contact, authorship, contestability, and continuity of judgment. That is the terrain on which alignment evidence either remains meaningful or begins to fail.

Acknowledgments: AGI-26 related drafting support and editorial iteration informed the preparation of this contribution.

Conflicts of Interest: The author has no competing interests to declare that are relevant to the content of this article.

References

1. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mane, D. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* **2016**.
2. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking: New York, 2019.
3. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models, 2021. arXiv preprint arXiv:2108.07258.
4. Goertzel, B. Judging the Journey, Not the Steps: Heavy Tails, Dam-Hard Problems and the Rationality of Collective Sacrifice, 2025. Preprint, v12 (October 4, 2025).
5. Goertzel, B. Economic Trajectories Toward Technological Singularity: Analysis via "Hyper-Intelligent Economics", 2025. Preprint, v2 (December 8, 2025).
6. Hyland, D.; Albarracin, M. On the Variational Costs of Changing Our Minds, 2025. arXiv preprint arXiv:2509.17957.
7. Albarracin, M.; Pitliya, R.J.; Smithe, T.S.C.; Friedman, D.A.; Friston, K.; Ramstead, M.J.D. Shared Protentions in Multi-Agent Active Inference. *Entropy* **2024**, *26*, 303. https://doi.org/10.3390/e26040303.
8. Hutchins, E. *Cognition in the Wild*; MIT Press: Cambridge, MA, 1995.
9. Clark, A.; Chalmers, D.J. The Extended Mind. *Analysis* **1998**, *58*, 7–19. https://doi.org/10.1093/analys/58.1.7.
10. Weick, K.E. *Sensemaking in Organizations*; Sage: Thousand Oaks, CA, 1995.
11. Montes, G.A. Causal Biomimesis: Self-Replication as Evolutionary Consequence. In *Biomimetic and Biohybrid Systems: 6th International Conference, Living Machines 2017, Stanford, CA, USA, July 26–28, 2017, Proceedings*; Mangan, M.; Cutkosky, M.; Mura, A.; Verschure, P.F.M.J.; Prescott, T.; Lepora, N., Eds.; Springer International Publishing: Cham, 2017; pp. 328–347. https://doi.org/10.1007/978-3-319-63537-8_28.

 doi:10.20944/preprints202604.2159.v1

12. Friston, K. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience* **2010**, *11*, 127–138. https://doi.org/10.1038/nrn2787.