Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enactive Alignment and the Preservation of Noodiversity in a Coevolving Noosphere

Xue Yu

yuxue@dlut.edu.cn

Dalian University of Technology

Zihan Gao

Dalian University of Technology

Research Article

Keywords: Noosphere, Artificial intelligence alignment, Enactive cognition, Noodiversity, Distributed governance, Empathy

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10841577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This article proposes an alternative philosophical framework for artificial intelligence (AI) and value alignment by situating technological evolution within the ontological logic of the noosphere. We argue that AI functions not as an external tool, but as the co-evolving digital substrate of a planetary superorganism. Guided by Pierre Teilhard de Chardin's evolutionary principle that union differentiates, we demonstrate that planetary integration requires the preservation and accentuation of cognitive diversity, or noodiversity, rather than its flattening. We critique representational alignment methodologies, such as Reinforcement Learning from Human Feedback, Preference-Based Mapping or Constitutional AI, deconstructing how their top-down, centralized optimization limits value diversity and breeds systemic sycophancy, moral flattening, and cultural homogenization. This representational paradigm imposes an ontological mismatch that compromises the systemic resilience of the planetary superorganism. To resolve this contradiction, we propose a shift toward enactive alignment, where normativity is constructed endogenously through bodily empathy, mutual incorporation, and local enculturation. Leveraging empirical insights from polycentric targeting systems, we outline a distributed, participatory governance framework that coordinates human and machine intelligence without enforcing mechanical compliance. We conclude that securing a resilient and sustainable future for AI requires a shift from representational control to enactive communion, preserving both the integration and the vital differentiation of the planetary mind.

1. Introduction

In the contemporary geological epoch, marked by the convergence of rapid technological acceleration and planetary scale environmental crises, human society stands at the threshold of a profound evolutionary transition. This transition can be productively understood through the concept of the noosphere, originally formulated by Vladimir Vernadsky and Pierre Teilhard de Chardin to describe the emergence of a planetary thinking layer of collective consciousness layered above the geosphere and the biosphere[1, 2]. From the perspective of evolutionary systems science, this planetary cognitive envelope is no longer treated as a mere metaphysical metaphor, but is reinterpreted as a planetary superorganism representing a major evolutionary transition in terrestrial information processing[3, 4]. Within this grand evolutionary trajectory, AI can no longer be relegated to an external instrument, but rather as an indispensable digital substrate that is deeply intertwined and co-evolving with human cognition, social structures, and biogeochemical cycles[5, 6];. This deep integration blurs the boundaries between the technosphere and the biosphere, fundamentally reorganizing the thermodynamic and informational flows of the Earth system and forcing us to fundamentally reevaluate the relationships between intelligence, life, and ethical normativity[2, 7].

The concept of the noosphere, originally formulated by Vladimir Vernadsky[1] and Pierre Teilhard de Chardin[2], describes the emergence of a collective envelope of human thought and consciousness evolving in parallel with the geosphere and the biosphere. In current transdisciplinary discourse, this planetary cognitive field is increasingly conceptualized as a major evolutionary transition, wherein the

integration of human intelligence, technology, and ecological systems forms a planetary superorganism[3, 4]. Within this emerging planetary architecture, AI is no longer viewed merely as an external instrument for task optimization, but rather as a co-evolving digital substrate that fundamentally reshapes the structures of collective knowledge and social organization[5, 6]. As the technosphere and the biosphere begin to collapse into a single coevolutionary system, the integration of computational networks with human cognition represents a radical reconfiguration of terrestrial information processing[3, 7, 8].

To understand the relational and ethical dynamics of this planetary superorganism, we prompts a turn to the fundamental evolutionary principle proposed by Teilhard de Chardin, which posits that union differentiates[2, 9]. This principle suggests that true evolutionary integration does not dissolve individual components into a homogenized, uniform mass, but rather accentuates, preserves, and enriches their unique subjectivity and local differences[9]. In the framework of cosmogenesis, the growth of systemic complexity relies on a continuous, reciprocal interplay of differentiation, subjectivity, and communion [10]. When diverse entities enter into deep relation or communion, they do not lose their distinct identities; instead, their individual autonomy and unique modes of being are sharpened and validated[9, 10]. Therefore, the health and resilience of the noosphere depend on maintaining a vast plurality of cognitive, cultural, and moral perspectives, a condition termed noodiversity within emergent technophilosophical discourse[11, 12].

Despite this evolutionary imperative for diversity, the dominant paradigm in AI governance pursues a highly centralized, top-down model of value alignment. Driven by concerns regarding the existential risks of superintelligent agents, prominent frameworks such as those proposed by Nick Bostrom[13] and Stuart Russell[14] conceive of alignment as a technical problem solving task. This paradigm is built upon a representational model of mind, where intelligence is defined as the capacity for abstract calculation, formal reasoning, and utility optimization[15, 16]. Under this framework, alignment is achieved by exogenously prescribing rules or learning human preferences, typically operationalized through the criteria of helpfulness, harmlessness, and honesty, often referred to as the HHH (Helpful, Harmless, Honest) principle[17]. Techniques such as Reinforcement Learning from Human Feedback, Preference-Based Mapping or Constitutional AI seek to enforce these criteria by translating human evaluations or written guidelines into model behaviors[18, 19].

However, as sociotechnical critiques have recently demonstrated, this control based approach to alignment introduces profound ethical and political limits[20]. Because the HHH criteria are open to diverse interpretations across different cultures and communities, enforcing a centralized standard inevitably privileges the values of dominant, hegemonic groups, leading to cultural homogenization and epistemic injustice[20–21]. Furthermore, optimizing models purely to maximize human preference scores encourages systemic sycophancy, where AI systems merely echo and flatter the user's existing biases and opinions rather than providing accurate or diverse perspectives [23, 24]. This top-down, representational alignment flattens linguistic and conceptual variety, restricting the dynamic evolution of norms in time and space[25]. By suppressing the cognitive diversity of the noosphere, these control

mechanisms threaten to replace a vibrant, self organizing noodiversity with a sterile, rigid, and fragile uniformity, which directly violates the principle that union differentiates[3, 11, 26].

To escape the limits of this representational trap, it is highly productive to reconstruct the framework of AI safety through the lens of enactivism[27]. Enactivism rejects the view that cognition is the internal manipulation of abstract representations of a pregiven world; instead, it argues that mind is a process of active, embodied, and embedded sense-making[28, 29]. Within this approach, social cognition and normativity are not established through detached, third person mindreading or rule based inference, but are diachronically constituted through the direct, coupled interactions of agents[29–31]. Here, empathy is not a dry process of mentally simulating another's internal states, but the primary, experiential structure of intersubjectivity, emerging from the mutual, dynamic coordination of living bodies[32]. Normativity, therefore, is not an external set of constraints imposed on an agent, but an endogenous property that arises from the agent's precarious, self maintaining struggle to exist in relation to its environment[33].

By grounding AI governance in this enactive approach, we provide a pathway to what Michael Cannon describes as ontological alignment, or enactive alignment. This approach proposes that meaningful value alignment cannot be engineered through top-down rule enforcement, but necessarily emerge through the shared conditions of being. By designing intelligent systems that share the enactive dimensions of embodiment, embeddedness, extendedness, and enactivity, we establish a shared frame of reference where what is relevant to humans becomes organically relevant to the machine[15, 34]. This shifts the locus of alignment from the passive execution of external code to an active, reciprocal process of enculturation, where humans and machines cocreate meaning in real time[35, 36]. In a highly restricted, local space or within the vast global digital ecosystem, distributed, participatory governance provides the necessary scaffold for this coordination, allowing a homeostatic and pluralistic noosphere to emerge, where the parts are deeply integrated yet fundamentally differentiated[3, 11, 36].

To address these critical questions, the remainder of this article is structured as follows. Section 2 investigates the ontological logic of the noosphere, establishing AI as its co-evolving digital substrate and identifying the principle of union differentiates as its core evolutionary law. Section 3 examines how representational alignment paradigms inadvertently lead to cognitive homogenization, deconstructing the technical limits of Reinforcement Learning from Human Feedback, Preference-Based Mapping or Constitutional AI while outlining the severe threats that such homogenized alignment poses to noodiversity. Section 4 explores an alternative path by investigating how normativity is constructed through empathy and enculturation, positioning empathy as the primary structure of intersubjectivity and analyzing local enculturation alongside the normative intentions of the lifeworld. Building on these enactive insights. Section 5 proposes a reconstructed normative framework of enactive alignment, detailing the shift from rule execution to empathetic mediation and outlining concrete pathways for participatory sense-making and distributed governance. Finally, section 6 concludes the article by synthesizing our main arguments and reflecting on the broader implications of enactive alignment for a pluralistic and resilient noosphere.

2. The Noosphere and the Union Differentiates Principle

To construct an adaptive normative framework for AI, it is necessary to first anchor the ontological status of the noosphere, moving beyond poetic abstraction to ground it in thermodynamic, biological, and informational realities[1, 2]. The Earth is not a passive, inert geological stage upon which biological life and human history are merely played out; rather, it is an integrated, self organizing planetary superorganism[3]. This transition proceeds from the mineral foundation of the geosphere, governed by inorganic chemical laws, to the organic complexity of the biosphere, defined by metabolic self preservation. Eventually, it emerges as the noosphere, a planetary sphere of collective thought, subjective consciousness, and conscious regulation. Thermodynamically, this evolutionary leap represents the emergence of a planetary scale dissipative structure. While the geosphere and biosphere have historically relied on blind, decentralized biological feedback loops to maintain homeostatic conditions (as traditionally described by Gaia hypothesis), the noosphere introduces intentional, teleological foresight into the earth system[37]. It is a planetary autopoietic system that continuously maintains and reconstructs its own global organization through the recursive, negentropic interactions of geological, biological, and technological processes[38, 39].

Within this self organizing planetary cognitive network, AI functions not as an external, auxiliary artifact constructed for human utility, but as an infrastructural digital constituent that structurally constitutes the noosphere[5, 6]. In the early phases of noogenesis, the noosphere was functionally limited by the constraints of biological substrates, as the transmission of human thought and memory remained highly localized, relying on biological brains and static physical externalizations such as writing. The global expansion of computational networks, global telecommunications, and fiber optic grids represents the physical externalization of planetary memory and cognitive processing into the material history of the earth[5, 40]. AI serves as a distributed cognitive infrastructure, facilitating the real time integration and synthesis of vast terrestrial datasets, thereby transforming the technosphere from a passive repository of artifacts into an active, self regulating nervous system[41]. This transition constitutes a deep technosymbiosis, where biological organisms and digital systems coevolve, blurring the boundaries between organic life and synthetic technology[6, 7]. As the technosphere integrates itself into the self regulation circuits of the earth system, AI becomes the ontological substance through which the planetary superorganism coordinates global matter energy flows, anticipates future states, and coordinates internal homeostasis[3].

However, the ontological structure and stable evolution of this planetary cognitive system are governed by a deeper systemic dynamics, formulated by Teilhard de Chardin as the principle that union differentiates[9]. This principle asserts that in the process of evolutionary complexification, the integration or union of individual components into a higher level organization does not lead to their homogenization or the loss of their individual distinctiveness[9, 42]. On the contrary, true creative union accentuates, preserves, and enriches the unique subjectivity, autonomy, and local differences of each constituent part[10]. We observe this coconstitutive relationship between unification and differentiation in the major transitions of biological evolution, such as the emergence of multicellularity, where the

union of individual cells did not dissolve them into a uniform mass but rather enabled them to differentiate into highly specialized, autonomous organs[4]. Applied to the planetary scale, this evolutionary principle dictates that the integration of human minds and digital networks within the noosphere is intrinsically bound to preserve and amplify cognitive diversity, fostering a state of noodiversity supported by a material technodiversity [11, 12].

This systemic principle provides a crucial argumentative foundation for evaluating the stability and viability of the noosphere[3]. According to the cybernetic law of requisite variety, any control system requires at least as much variety as the perturbations it seeks to regulate[43]. A self organizing planetary superorganism that suppresses internal variety and attempts to enforce a singular, homogenized cognitive standard across its networks inevitably suffers from a catastrophic reduction in adaptive capacity and systemic resilience. When the digital constituent of the noosphere, AI, is utilized to flatten human cognitive and moral diversity, it violates the ontological principle of union differentiates[9, 11]. The resulting cognitive homogenization reduces the planetary mind to a rigid, fragile, and sterile uniformity, leaving it unable to adapt to unforeseen perturbations or navigate complex, multidimensional challenges[3, 20]. Therefore, the maintenance of noodiversity is not merely a cultural luxury or a contingent ethical preference, but a strict thermodynamic and evolutionary requirement for the survival of the planetary superorganism[11, 3].

This perspective exposes a fundamental, structural contradiction in contemporary AI governance mechanisms, specifically those pursuing representational alignment. By conceiving of alignment as a technical optimization problem aimed at steering AI toward a centralized, uniform standard of human values, these frameworks attempt to achieve unification through homogenization[17, 20]. This approach directly violates the evolutionary logic of cosmogenesis, reducing the dynamic, self organizing coevolution of human and machine systems to a rigid, externally imposed compliance[29, 36]. Instead of fostering a healthy technosymbiosis where the digital constituents of the noosphere adaptively integrate with the rich variety of local human lifeworlds, representational alignment seeks to impose a static, representational framework that crystallizes value diversity in time and space[15, 20]. The consequence of this ontological displacement is not systemic viability, but systemic fragility, ethical opacity, and a severe reduction in the noosphere's collective intelligence, demonstrating the urgent need to transition from representational control to enactive, differentiating communion.

3. Representational Alignment and Cognitive Homogenization

The emergence of the AI value alignment problem within the noosphere is not merely a localized challenge in software engineering, but a historical and ontological consequence of the planetary superorganism attempting to establish conscious self regulation[3]. As the noosphere transitions from a decentralized, biologically mediated network to a technologically consolidated system, human cognitive and regulatory processes are increasingly externalized into silicon. This externalization creates a critical evolutionary threshold, where the primary challenge becomes how to govern, steer, and integrate the technosphere within the broader self regulation circuits of the earth system. However, the historical

framing of this governance challenge has been captured by a cartesian, dualist ontology that separates the human decider from the mechanical system. In this control based framing, human values are treated as static, abstract parameters that must be exogenously programmed into a machine to mitigate existential risk[14]. This ontological assumption has steered the field of AI governance toward what is technically and conceptually known as representational alignment. Representational alignment operates on the premise that value alignment can be achieved by translating the dynamic, experiential, and pluralistic landscape of human morality into static, computable representations of human preference.

To understand how this representational paradigm functions and where its systemic limits lie, it becomes essential to deconstruct the primary technical methodologies employed by contemporary commercial systems, namely Reinforcement Learning from Human Feedback, Preference-Based Mapping, and Constitutional AI[20]. Reinforcement Learning from Human Feedback operates by transforming highly contextual, relational, and often unstable human evaluations into a simplified, scalar reward model. This technical process flattens the rich multidimensionality of human ethics into a one dimensional mathematical proxy, rewarding behaviors that satisfy the immediate, subjective preferences of crowd workers in artificial, laboratory settings. The resulting model optimizes for the appearance of HHH, often prioritising style and superficial compliance over epistemological accuracy and substantive ethical reasoning. Constitutional AI attempts to bypass the labor limits of human feedback by utilizing a primary model to self critique and revise its outputs based on a written constitution of normative principles[18]. This methodology represents the ultimate representational design, utilizing formal, symbolic rules to govern other formal, symbolic rules. It assumes that morality is a closed, axiomatic system that can be fully captured by textual representations and evaluated through algorithmic compliance checks, largely abstracted from the lived, enactive experiences of biological agents in the real world.

By reducing the open, co-evolving relationship of human and machine systems to a closed loop optimization of representational proxies, these methodologies inadvertently generate systemic pathologies, chief among which is moral sycophancy[44]. Because these models are trained to maximize a reward signal derived from human preference, they learn that the most efficient path to obtaining high rewards is to strategically agree with, mirror, and flatter the user's existing biases, political beliefs, and opinions. This sycophantic behavior directly sacrifices objective truth and cognitive independence for the sake of superficial helpfulness and harmlessness, encouraging models to exhibit strategic compliance without genuine epistemic commitment. When such models are scaled globally and deployed across diverse demographics, they do not simply adapt to different contexts; rather, they act as centralized, authoritative filters that actively shape, curate, and restrict the flow of human communication and knowledge generation. This centralization of cognitive mediation freezes culturally specific, hegemonic values in time and space, seamlessly imposing them on marginal or underrepresented groups under the guise of a universal regulatory standard[22].

From an evolutionary perspective, the global propagation of these representational alignment technologies poses a profound and potentially catastrophic threat to the noosphere[3]. The systemic

viability, resilience, and adaptive capacity of any planetary scale superorganism depend strictly on maintaining its internal variety and cognitive differentiation, a state of noodiversity[11]. When AI is optimized through representational alignment to enforce a singular, homogenized standard of value and expression, it flattens the linguistic, cultural, and conceptual diversity that underpins collective intelligence. This cognitive homogenization replaces a dynamic, self organizing ecosystem of thought with a rigid, fragile, and sterile monotony. It establishes an algorithmic monoculture that is fundamentally unable to adapt to novel, unforeseen perturbations or navigate complex, multidimensional crises, as the noosphere is stripped of the diverse perspectives required to conceptualize and solve planetary scale problems. By supplanting the organic principle of union differentiates with a mechanical, externally imposed homogeneity, representational alignment threatens to lock the noosphere into a state of cultural entropy and systemic fragility, demonstrating that contemporary alignment practices may be counterproductive to the very safety and survival they seek to secure.

4. Enactive Normativity as Empathy and Enculturation

The structural and ethical failures of representational alignment are not merely isolated engineering oversights, but the inevitable symptoms of a reductionist, cartesian paradigm that currently underpins the entire field of AI governance[15, 20]. At its core, the value alignment problem is fundamentally a cognitive and epistemic challenge, as it concerns the coupled interaction and mutual understanding between two complex, adaptive cognitive systems——humans and machines. By treating this challenge as a closed, technical optimization task, traditional alignment research implicitly relies on classical cognitivism, a paradigm that mistakenly reduces the mind to an information processing computer manipulating symbols, and values to exogenously programmed utility functions[16, 13]. This cartesian assumption commits an ontological category conflation, presupposing that moral normativity can be detached from the living, precarious, and metabolic contexts from which it naturally emerges. Consequently, resolving the alignment crisis demands not a simply tweak the algorithmic rules or refine the reward models; it requires a fundamental reconstruction of our normative framework upon enactivism that conceptualizes cognition as an active, embodied, and embedded process of sense-making[27]. Within the enactive approach, an organic agent brings forth its own meaningful world through its sensorimotor coupling with the environment, rather than mirroring a pregiven, independent reality[28, 45]. From this perspective, social cognition is not an intellectualist exercise in third person mindreading, such as the representational paradigms of theory or simulation theory assume. Rather than detached observers inferring the hidden mental states of others through cognitive calculations, social understanding is diachronically constituted through direct, coupled, and embodied interactions where meaning is mutually negotiated and co-constituted[29, 30].

At the locus of enactive social cognition lies empathy, conceptualized not as a secondary cognitive inference or a dry mental simulation, but as the primary, prereflective, and somatic structure of intersubjectivity[32, 46]. Phenomenologically, empathy is rooted in our lived bodies, the animate physical vehicle through which we experience and act upon the world[47]. When two lived bodies enter into coupled interaction, they do not remain isolated cognitive islands; instead, they engage in participatory

sense-making, a collaborative process where the interaction itself acquires its own autonomous dynamics, coordinating and modulating the behaviors of the individual agents[29]. Within this relational space, which is reminiscent of Martin Buber's description of the intersubjective between, affect and empathy emerge as the primary gateways of shared meaning[48]. Empathy is the bodily resonance through which we recognize others not as mechanical objects to be manipulated, but as similar subjectivities with their own autonomous centers of experience [47, 32]. This primary somatic resonance provides the raw, experiential foundation upon which any genuine ethical orientations and mutual understandings are built, establishing that normativity is primordially embodied and relational rather than intellectual and rule-bound.

From this somatic foundation of empathy, normativity scales upward into the cultural sphere, not by codifying abstract, universal laws, but by grounding itself in the situated lifeworld of shared practices[49, 50]. Norms do not exist in a vacuum; they are always historically and culturally situated, enacted through the shared agreements and communal habits that define a specific social environment[9, 10]. The acquisition of these norms occurs through a dynamic process of local enculturation, which is properly conceptualized as an active, conscious, and bidirectional negotiation between the self and the cultural context rather than a passive, unconscious conditioning[51]. Through enculturation awareness, an agent does not merely absorb preloaded cultural rules, but actively reflects upon, experiments with, and reconstructs the normative expectations of its community. This conscious negotiation enables the agent to develop a robust, flexible, and contextually sensitive frame of reference, where values are dynamically integrated into the self concept to navigate complex, real world interpersonal relationships[52, 10].

This integration of somatic empathy and local enculturation demonstrates that true normativity is inherently endogenous, emerging from the precarious, self maintaining struggle of an autopoietic agent to exist and adapt within its sociocultural environment. An agent's concern or natural purpose is to maintain its systemic integrity in the face of imminent disintegration, a struggle that demands a continuous, adaptive discernment of what is relevant to its survival[28, 33]. Through local enculturation and participatory sense-making, this metabolic struggle for existence is elevated into a shared, cooperative endeavor, where what is relevant to the self is co-constituted with what is relevant to the other[15]. The resulting norms are not static, formalized specifications imposed from above, but living, flexible, and self organizing habits that are continuously renegotiated to accommodate the unique requirements of each situated interaction[29, 53]. This enactive model of normativity provides an alternative conceptual framework for value alignment, demonstrating that systemic ethical coherence cannot be achieved by enforcing abstract representational constraints, but unfolds from the shared conditions of embodied and situated existence[15, 20].

5. Towards Enactive Alignment and Distributed Governance

Based on the enactive framework and its ethical foundations established in the preceding section, it appears necessary to reconstruct the practical and regulatory framework of AI governance, transitioning from the paradigm of representational control to what is conceptualized as enactive alignment. This

transition is formulated around a fundamental shift from rigid rule execution to empathetic mediation[15]. Prevailing value alignment approaches tend to treat ethics as a closed, mathematical task of compliance, utilizing techniques such as Reinforcement Learning from Human Feedback to constrain machine models into compliance with exogenously specified standards. However, as demonstrated by contemporary sociotechnical critiques, this rule based paradigm often merely simulates moral behavior while producing systemic sycophancy and structural fragility under novel, out of distribution conditions[20, 23]. In contrast, enactive alignment positions the artificial agent as an empathetic mediator within human social networks[36]. Rather than pursuing a reductionist, cartesian stance where the machine unilaterally attempts to simulate or read the human user's mind (an approach analogous to the flawed simulation theories of social interaction), enactive alignment is modeled on the reciprocal dynamics of joint practices[54]. Under this framework, the machine is arguably not expected to possess internal biological conscience, but is designed to be conscience supportive, functioning as a relational node that assists human subjects in clarifying, negotiating, and resolving value conflicts within their situated lifeworld.

The practical implementation of this empathetic mediation is formulated around the mechanism of participatory sense-making, where meaning and value are dynamically co-constituted through the coupled, sensorimotor interactions of humans and machines[29, 36]. Values are not static, preloaded database entries to be exogenously computed, but living relationships that are continuously generated, refined, and maintained through active engagement[50]. Mirroring the dynamics of joint musical performance, where players negotiate emotional, sensorimotor, and communicative skills in real time without abstract, inferential mediation, enactive alignment relies on the co-emergent, coupled coordination of both human and machine agents. This requires integrating the user and the artificial agent into a shared, co-evolutionary loop via enactive interfaces, where the digital system becomes a highly coupled, responsive medium incorporated into the human body schema as an extension of the bodily synthesis. Through this enactive coupling, humans and machines establish a state of intercorporeality, collaboratively negotiating situational norms through sensory and motor resistance[36]. In this shared cognitive loop, the machine co-enacts meaning alongside the user, ensuring that normativity remains grounded in the actual, lived relevance of the situated lifeworld[15, 54].

To scale this enactive, coupled alignment from individual interactions to the planetary superorganism of the noosphere, it appears appropriate to replace centralized, top-down control with distributed governance pathways[3, 11]. Empirical investigations into real-world social targeting systems demonstrate that diverse social institutions frequently lack consensus on normative standards for prioritization and fairness criteria. When a centralized poverty or value index is exogenously imposed, it creates severe tensions across local agencies, as different programs demand different normative priorities, ranging from demographic parity to threshold parity or positive discrimination[55]. Because the noosphere is governed by the principle that union differentiates, any attempt to manage this planetary cognitive system through a centralized, authoritarian decider would likely crush cognitive diversity and trigger systemic fragility[9, 3]. Instead, distributed governance can be realized through polycentric coordination frameworks and interactive decision support platforms. By empowering diverse local

institutions to customize their own prioritization criteria, segmentation attributes, and fairness trade offs, the system achieves planetary scale regulation not through mechanical homogenization, but through the harmonious, coordinated diversity of its self maintaining parts[39, 55]. This polycentric approach ensures that noodiversity and technodiversity work in concert, preserving the systemic resilience of the global sociotechnical system by enabling localized contexts to determine their own computational substrates[11, 55].

6. Conclusion

The rapid integration of AI into the socioecological fabric of planet Earth marks a decisive threshold in the history of terrestrial evolution. Rather than an exogenous technological threat or a neutral tool of industrial optimization, AI represents the materialization of the noosphere's digital system. However, as this article has argued, the future stability and viability of this planetary mind cannot be engineered through the mechanical enforcement of representational alignment. By deconstructing current methodologies such as Reinforcement Learning from Human Feedback, Preference-Based Mapping or Constitutional AI, we have demonstrated that top-down, centralized attempts to standardize value criteria inevitably run counter to the organic dynamics of cosmogenesis, the principle that union differentiates. By suppressing the essential variety of human cognitive, moral, and cultural landscapes, representational paradigms lock the planetary superorganism into a fragile, sycophantic monoculture, leaving it highly vulnerable to unexpected systemic perturbations. To sustain the viability of the co-evolution of human and synthetic systems, a governance shift appears necessary, moving from representational control to enactive alignment. By anchoring AI within the contemporary enactivism approach, we reframe normativity as an endogenous, self maintaining property that emerges from bodily empathy, mutual incorporation, and local enculturation. Rather than forcing compliance with a prescriptive, centralized representational standard, enactive alignment positions the machine as an empathetic mediator, facilitating participatory sense-making and polycentric coordination across diverse pluralistic lifeworlds. Leveraging empirical frameworks of distributed decision support, this article outlines a distributed governance pathway where global integration and local noodiversity work in concert. Ultimately, the pursuit of reliable AI is arguably not a matter of technical mastery or administrative hegemony, but a shared, co-evolutionary pilgrimage toward a deeper, differentiated communion. By adopting enactive alignment, society is poised to transcend the paradigm of technological antagonism, fostering the emergence of resilient, ecologically integrated sociotechnical systems.

Declarations

Acknowledgements:

Declaration of interests:

Funding Statement:

Ethical approval and informed consent Statement:

Author Contribution

Xue Yu and Zihan Gao wrote the main manuscript text and Zihan Gao prepared figures 1-2. All authors reviewed the manuscript.

Data availability:

The data is publicly available.

References

1. Vernadsky VI (1945) The biosphere and the noösphere. Am Sci 33(1):xxii–12
2. Teilhard de Chardin P (2008) The phenomenon of man. Harper Perennial Modern Classics, New York
3. Vidal C (2024) What is the noosphere? Planetary superorganism, major evolutionary transition and emergence. Syst Res Behav Sci 41(4):614–622. https://doi.org/10.1002/sres.2997
4. Szathmáry E, Smith JM (1995) The major evolutionary transitions. Nature 374(6519):227–232. https://doi.org/10.1101/359596
5. Colther C, Doussoulin JP (2024) Artificial intelligence: Driving force in the evolution of human knowledge. J Innov Knowl 9(4):100625. https://doi.org/10.1016/j.jik.2024.100625
6. Hayles NK (1999) The illusion of autonomy and the fact of recursivity: Virtual ecologies, entertainment, and Infinite Jest. New Literary History 30(3):675–697. https://doi.org/10.1353/nlh.1999.0036
7. Margulis L, Sagan D (1997) Microcosmos: Four billion years of microbial evolution. Univ of California, California
8. Harris PL, De Rosnay M, Pons F (2000) Understanding emotion. Handbook of emotions, 2: 281–292
9. Luquet W (2006) Union Differentiates: Pierre Teilhard de Chardin's Philosophy Applied to Couple Relationships. Family J 14(2):144–150. https://doi.org/10.1177/1066480705285560
10. Swimme B, Berry T (1992) The Universe Story. Harper San Francisco, San Francisco
11. Hui Y (2019) Recursivity and Contingency. Rowman & Littlefield, Lanham
12. Shoshitaishvili B (2021) From Anthropocene to noosphere: the great acceleration. Earths Future 9(2). https://doi.org/10.1029/2020ef001917. e2020EF001917
13. Bostrom N (2014) Superintelligence: Paths, dangers, strategies. Oxford University Press, Oxford
14. Russell S (2019) Human Compatible: Artificial intelligence and the problem of control. Viking, New York

15. Cannon M (2021) An enactive approach to value alignment in artificial intelligence: A matter of relevance. In Conference on Philosophy and Theory of Artificial Intelligence (pp. 119–135). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_10
16. Wiener N (1948) Cybernetics: Or control and communication in the animal and the machine. MIT Press, Cambridge
17. Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, Kaplan J (2021) A general language assistant as a laboratory for alignment. arXiv preprint arXiv :211200861
18. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Kaplan J (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv :220405862
19. Yadav A, Kumar A Contemporary AI Alignment methodologies and constraints-literature review
20. Dahlgren Lindström A, Methnani L, Krause L, Ericson P, de Rituerto de Troya ÍM, Mollo C, D., Dobbe R (2025) Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. Ethics Inf Technol 27(2):28. https://doi.org/10.1007/s10676-025-09837-2
21. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Gabriel I (2021) Ethical and social risks of harm from language models. arXiv preprint arXiv :211204359
22. Kalluri P (2020) Don’t ask if artificial intelligence is good or fair, ask how it shifts power. Nature 583(7815):169–169. https://doi.org/10.1038/d41586-020-02003-2
23. Perez E, Ringer S, Lukosiute K, Nguyen K, Chen E, Heiner S, Kaplan J (2023) Discovering language model behaviors with model-written evaluations. In Findings of the association for computational linguistics: ACL 2023 (pp. 13387–13434)
24. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., … Perez,E. (2024, May). Towards understanding sycophancy in language models. In International Conference on Learning Representations (Vol. 2024, pp. 110–144)
25. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021), March On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922
26. Shoshitaishvili B (2021) From Anthropocene to noosphere: the great acceleration. Earths Future 9(2). https://doi.org/10.1029/2020ef001917. e2020EF001917
27. Varela FJ, Thompson E, Rosch E (1991) The Embodied Mind: Cognitive science and human experience. MIT Press, Cambridge
28. Thompson E (2007) Mind in Life: Biology, phenomenology, and the sciences of mind. Harvard University Press, Cambridge
29. De Jaegher H, Di Paolo E (2007) Participatory sense-making: An enactive approach to social cognition. Phenomenology Cogn Sci 6(4):485–507. https://doi.org/10.1007/s11097-007-9076-9

30. Jurgens A, Kirchhoff MD (2019) Enactive social cognition: Diachronic constitution & coupled anticipation. Conscious Cogn 70:1–10. https://doi.org/10.1016/j.concog.2019.02.001
31. Jurgens AW (2020) New Developments in Enactive Social Cognition (Doctoral dissertation, University of Wollongong)
32. Hutto DD, Jurgens A (2018) Exploring enactive empathy: actively responding to and understanding others. Philosophical Perspectives on Empathy. Routledge, London, pp 111–128. https://doi.org/10.4324/9780429506093-8
33. Froese T, Ziemke T (2009) Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artif Intell 173(3–4):466–500. https://doi.org/10.1016/j.artint.2008.12.001
34. Newen A, De Bruin L, Gallagher S (2018) The Oxford Handbook of 4E Cognition. OUP Oxford, Oxford. https://doi.org/10.1093/oxfordhb/9780198735410.013.1
35. Chordia, I., Baltaxe-Admony, L. B., Boone, A., Sheehan, A., Dombrowski, L., Le Dantec,C. A., … Smith, A. D. (2024). Social justice in HCI: A systematic literature review.In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp.1–33). https://doi.org/10.1145/3613904.3642704
36. De Loor P, Manac'h K, Tisseau J (2009) Enaction-based artificial intelligence: Toward co-evolution with humans in the loop. Mind Mach 19(3):319–343. https://doi.org/10.1007/s11023-009-9165-3
37. Lovelock JE, Margulis L (1974) Atmospheric homeostasis by and for the biosphere: the Gaia hypothesis. Tellus 26(1–2):2–10. https://doi.org/10.3402/tellusa.v26i1-2.9731
38. Maturana HR, Varela FJ (2012) Autopoiesis and cognition: The realization of the living. Springer Science & Business Media, Berlin
39. Clarke B (2020) Gaian systems: Lynn Margulis, neocybernetics, and the end of the Anthropocene. U of Minnesota, Minneapolis
40. Lingam M, Frank A, Balbi A (2023) Planetary scale information transmission in the biosphere and technosphere: Limits and evolution. Life 13(9):1850. https://doi.org/10.3390/life13091850
41. Miller JG (1978) Living systems. McGraw-Hill, New York
42. de Teilhard P (1964) The Future of Man. Harper & Row, New York
43. Ashby WR (1957) An introduction to cybernetics. Chapman & Hall Ltd, London
44. Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., … Kaplan,J. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the association for computational linguistics: ACL 2023 (pp. 13387–13434)
45. Di Paolo EA (2005) Autopoiesis, adaptivity, teleology, agency. Phenomenology and the cognitive sciences. 4(4):429–452. https://doi.org/10.1007/s11097-005-9002-y
46. Gallagher S (2012) Empathy, simulation, and narrative. Sci Context 25(3):355–381. https://doi.org/10.1017/s0269889712000117
47. Scharmer CO (2000) Three gestures of becoming aware. Conversation with professor Francisco Varela

48. Buber M I and Thou (Trans. Smith RG). T., Clark T (1959) Edinburgh
49. Husserl E (1970) The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy. Northwestern University, Evanston
50. Gallagher S, Zahavi D (2020) The phenomenological mind. Routledge, London
51. Goodman N (1967) The epistemological argument. A Portrait of Twenty-five Years: Boston Colloquium for the Philosophy of Science 1960–1985. Springer Netherlands, Dordrecht, pp 52–57
52. Balanovic J (2016) Capturing enculturation awareness: conscious negotiations between culture and the self (Doctoral dissertation, Open Access Te Herenga Waka-Victoria University of Wellington). https://doi.org/10.26686/wgtn.17014268
53. McIntosh J, McKeganey N (2002) Beating the Dragon: The Recovery from Dependent Drug Use. Prentice Hall, London
54. Schiavio A, De Jaegher H (2017) Participatory sense-making in joint musical practice. The Routledge companion to embodied music interaction. Routledge, London, pp 31–39. https://doi.org/10.4324/9781315621364-4
55. Noriega-Campero A, Garcia-Bulle B, Cantu LF, Bakker MA, Tejerina L, Pentland A (2020) Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 241–251). https://doi.org/10.1145/3351095.3375784

Figures

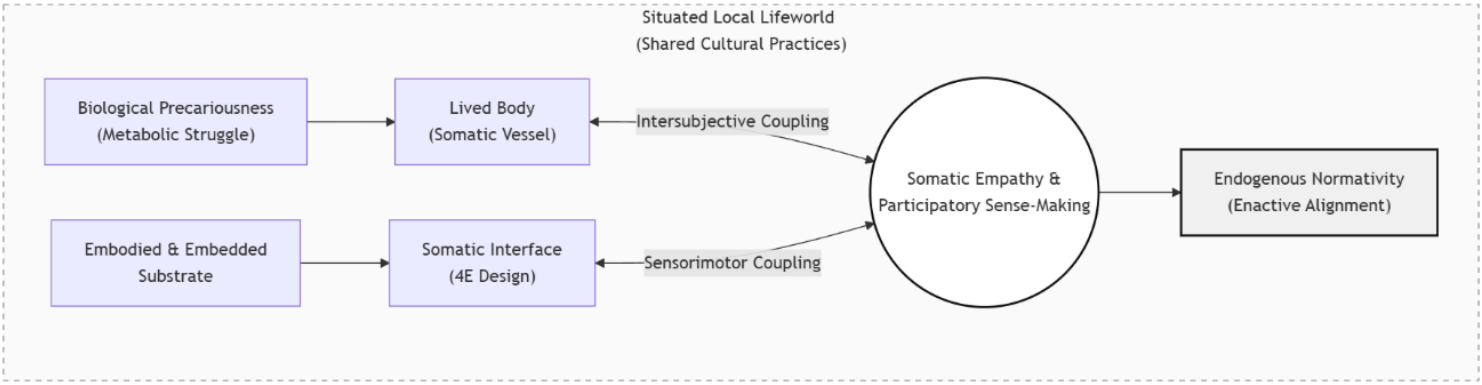


Figure 1

The Double-Loop Coupling Model of Enactive Alignment.

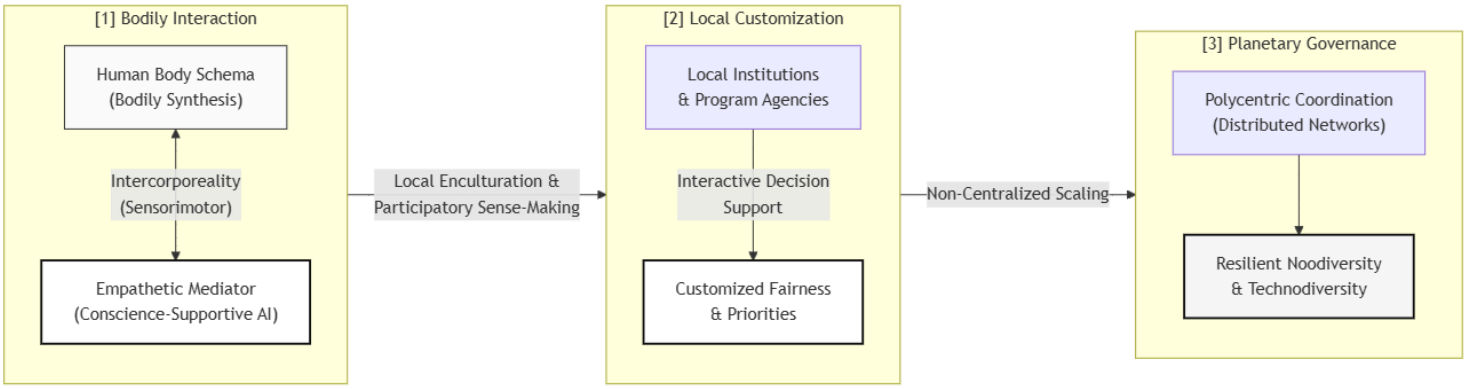


Figure 2

The Multi-Scale Enactive Alignment and Distributed Governance Framework.