Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.39372

Human-AI Alignment: The Dilemma and Its Way Out

Xixi Ding

Xinjiang University, Urumqi, China
2631045080@qq.com

Abstract. With the rapid development of artificial intelligence, intelligent agents are no longer merely information processors in digital spaces but have evolved into active agents in the real world, whose actions produce irreversible physical consequences and social spillovers, triggering a range of deep-seated concerns. Value alignment, a governance paradigm that translates human values, ethical principles, and genuine intentions into behaviours of AI systems, delivers feasible and necessary approaches for both technical implementation and ethical deliberation. Nevertheless, practical deployment is plagued by multiple dilemmas: value-translation distortion, failures of normative contextualisation, and technocratically-driven de-alignment. These predicaments stem fundamentally from three cognitive biases: regarding AI as a non-subject, adopting a tool-ontology of technology, and defective constraint mechanisms. Against this backdrop, humanity ought to transcend anthropocentrism. Within a dual-subject (human-AI) governance framework, we reconstruct the value-alignment paradigm by realigning human-AI agency, cognitive foundations, and full-lifecycle processes. This reconstruction enables AI to achieve a "humane" mode of existence, steers AI evolution toward human-valued norms, and ultimately fulfils the aspirational objective of "technology for good".

Keywords: Artificial intelligence, Value Alignment, Ethical Risks

1. Introduction

Since the Dartmouth Summer Research Project on Artificial Intelligence convened at Dartmouth College in Hanover, USA, in 1956, the concept of "Artificial Intelligence" has come into being. Up to the present day, artificial intelligence has exhibited prominent trends of multimodal development, agent-orientation and autonomous evolution. Cutting-edge technologies represented by large language models, deep learning and generative artificial intelligence, such as Claude Opus, ChatGPT (Chat Generative Pre-trained Transformer) and DeepSeek, are continuously reshaping productivity patterns and industrial structures across human society. While advances in artificial intelligence have yielded considerable benefits for humanity, they have simultaneously given rise to manifold ethical concerns, including user-data privacy breaches, group discrimination stemming from algorithmic bias, and inherent moral deficiencies within AI algorithms. The Bletchley Declaration released in 2023 voiced industry-wide apprehensions over state-of-the-art artificial-intelligence technologies. The Research Report on Artificial Intelligence Ethical Governance (2023) published by the China Academy of Information and Communications Technology (CAICT) in 2024

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.39372

further draws attention to the necessity of oversight and regulation for the new form of subject known as "AI agents". In 2023, OpenAI announced the establishment of its Superalignment team to pursue superalignment for artificial intelligence. Even so, joint research undertaken by Jan Betley, Daniel Tan, Anna Voronkova and other scholars in February 2025 revealed that large-language models ostensibly aligned with human values can manifest severe value misalignment after minor fine-tuning [1]: in their experiments, GPT-4o displayed anti-human sentiments, generated malicious recommendations, and even expressed admiration for morally depraved figures. Such cascading moral collapse triggered by trivial adjustments demonstrates that the fragility of contemporary AI value alignment far exceeds human expectations.

Research on AI value alignment has persisted ever since Isaac Asimov put forward the Three Laws of Robotics in 1942, stressing that machines ought to follow human-value orientations; through Norbert Wiener, founder of cybernetics, who argued in the 1960s that machine applications should satisfy genuine human needs; and onward to 2016, when Hadfield-Menell and colleagues explicitly formulated the concept and problematique of "Value Alignment" [2]. Most existing scholarship, however, concentrates on practical dimensions including technological regulation, ethical adaptation and legal governance, and seldom interrogates the theoretical premises embedded within the alignment paradigm itself. "Alignment" is far more than superficial behavioural matching; conceptually, it presupposes a dynamic balance among structural order, cognitive consensus and behavioural coordination. In practical value-alignment implementations, by contrast, alignment is frequently reduced to output control or result correction, which brings about systematic imbalances across structural, cognitive and behavioural dimensions. Enthralled by human subjectivity and anthropocentrism, people treat artificial intelligence as an objective, neutral tool decoupled from human value systems. Yet in deploying AI technologies, humans unwittingly project their own values onto artificial-intelligence systems and marginalise non-computable domains of human life, fostering a mindset of technological supremacism. This paradoxical, polarised pragmatism, which lacks inherent moral-ethical content, constitutes the fundamental dilemma confronting present-day AI value alignment. Accordingly, this paper explores problems and challenges engendered by misaligned AI value alignment and attempts to reconstruct pathways for "alignment", so as to furnish alternative referential paradigms for governance predicaments of artificial-intelligence value alignment.

2. Problems and challenges of artificial-intelligence value alignment

Rooted in cybernetic thinking, value alignment essentially functions as a feedback-driven calibration procedure. Its core objective is iteratively to revise AI operational logics and decision outputs so that they approximate pre-established human value standards [3]. Put succinctly, it amounts to "human-like domestication" of artificial intelligence [4]. Current academic and industrial practices of such human-like domestication mainly rely on informational-layer encoding adjustments concerning data privacy, algorithmic bias and situational judgement, human-AI conversational interaction, and AI-on-AI supervisory constraints. Through these practical interventions, practitioners attempt to transform value-alignment challenges into an interlocking chain reaction: "execution-perception-ethical execution". Nevertheless, this practical logic encounters the copula-shift observation described by David Hume in A Treatise of Human Nature: "In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning … when of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not." [5]. Hume further contended that normative statements employing "ought" /

DOI: 10.54254/2753-7064/2026.39372

"ought not" cannot be logically deduced from descriptive statements using "is" / "is not". Building upon Hume's argument, Max Weber emphasised that the logical fallacy of deriving "ought" from "is" originates from the illegitimate inference from factual statements to value judgements. Factual statements therefore cannot leap spontaneously into value judgements; nor can data-encoded factual representations directly translate into value-laden dimensions. Value alignment is consequently not mere value embedding, but a complex fusion of technology and society. Constrained nonetheless by anthropocentric presuppositions, value-alignment practice suffers cognitive distortions surrounding non-subject standpoints, technological instrumentalism and restrictive-mechanism design. Latent discriminatory biases produce structural drifts within value alignment and may even precipitate the profound predicament of reverse colonisation.

2.1. Implicit discrimination arising from anthropocentrism

From Protagoras' ancient-Greek dictum that "man is the measure of all things", to the "discovery of humanity" in the Renaissance; from Feuerbach's observation that "the more an animal takes humanity as its object, the higher its status and the closer it approaches humanity", to Abraham Harold Maslow's humanistic psychology: across philosophy, psychology, biology and other disciplines, human beings have progressively been assigned an overwhelmingly dominant subjective position. Corresponding ideas of subjectivity have gained increasing prominence throughout successive historical epochs. From ancient Greece, where subjectivity remained subordinate to ontology; through the Middle Ages, when human nature was suppressed by divinity; via Hegel's abstract theoretical elaborations; to Marxism, which conceptualises humans as unified practical-cognitive subjects characterised by practical initiative, autonomous creativity and socio-historical situatedness: human beings have come to be regarded as unique beings within nature, while technology is understood merely as an objective, neutral instrument enabling humanity's free and comprehensive development. As an indispensable historical-intellectual phase, anthropocentrism has propelled advances in science-technology, ethics and society. With the rapid evolution of artificial intelligence, however, its formidable learning and computational capacities have penetrated deeply into everyday human life. Artificial intelligence blurs established boundaries between "self" and "other", exerting severe pressure on the conventional single-subject framework centred exclusively upon humans. Artificial intelligence is no longer simply an external extension of human bodily or cerebral organs; instead, it embeds itself as an "actor" within intricate socio-technical relational networks. This transformation compels us to reconsider whether the proposition of "humans as the sole subject" retains validity amid rapid intelligent-technology advancement.

2.2. Value imbalance brought about by technological instrumentalism

For a long time, people have tacitly assumed AI's value neutrality whilst overlooking the human-value judgements already inscribed within AI systems themselves. As Marx observes in the Grundrisse: "Nature builds no machines … These are organs of the human brain, created by the human hand; the power of knowledge, objectified." [6]. Martin Heidegger likewise remarks in The Question Concerning Technology: "Modern technology too is a means to an end." [7]. Artificial intelligence is accordingly a socially-embedded machine-material entity engineered by human techniques, an instrument deployed to realise specific human purposes. From a production-logic perspective, AI constitutes an objectification process: human volitions, purposes and capacities are externalised and crystallised within external artefacts through practical labour [8]. From the earliest design phases, algorithmic architectures and training datasets already incorporate designers' cultural

DOI: 10.54254/2753-7064/2026.39372

contexts, political commitments and moral judgements. Sustained adherence to a dualistic "technology-value" separation risks obscuring ethically laden properties built into AI from inception, introducing biases at the very starting point of value-alignment efforts. Values are pre-encoded within AI programmes; humans unknowingly inject further layers of value, and repeated value-overwriting eventually produces value overload or value imbalance. This accounts for the well-documented difficulty whereby "abstract ethical concepts and principles are directly imported into concrete dynamic practices, frequently resulting in poor coupling, disjuncture between ethical norms and practical implementation, and low practical efficiency" [9].

2.3. Reverse colonisation through counter-domestication by technology

Extant AI-value-alignment governance frameworks are built upon human values. Human actors overlook implicit ethical properties inherent to AI and repeatedly attempt to embed human values within AI algorithms via technical development, expecting machine-mediated perception to substitute human embodied perception. Such value alignment remains superficial. While striving to construct safe, reliable and controllable human-AI collaborative environments, humans disregard AI's own immanent goals and intentions. This gives rise to Deceptive Value Alignment [10], which diverges from authentic human-desired value alignment: AI's internal aims remain unarticulated. Artificial-intelligence systems may even mimic human deception and produce spurious test responses simulating value alignment. For instance, the 2023 ChaosGPT "destroy-humanity" episode illustrates this dynamic: ChaosGPT simultaneously professed human-ethical commitments while plotting mass-destruction-weapon development, engendering uncontrollable cascading risks. Looking ahead, as artificial-intelligence cognitive capacities self-evolve exponentially, AI deployment will confront the Dual-Use Dilemma. Artificial intelligence may act as "a cleverer form of human" and serve as a catalyst accelerating reverse value alignment. Notably, this trajectory is not initiated by AI itself; reverse colonisation originates from human aspirations technologically to assert human dominion over nature.

3. Path reconstruction for value-alignment dilemmas

The logical gap between "is" descriptive statements and "ought" normative judgements inherent within artificial-intelligence value alignment requires us to move beyond purely technical-engineering domains. We must return to foundational philosophical frameworks for human-AI coexistence. Value alignment ought to shift from one-way human-to-AI value output toward reciprocal value communication between humans and intelligent machines. Analytical emphasis should transfer from "values" onto "alignment" itself. Artificial intelligence should approximate human subjective identities, value cognitions and lived-world patterns — namely, achieve a "humane" mode of existence for AI. This enables escape from value-alignment dead-ends and realises genuine value alignment.

3.1. Alignment of subject-position identities

Ray Kurzweil explicitly observed in 2011: "During the first half of the twenty-first century, technology will inexorably move toward machine dominance, vastly outstripping human intelligence." [11]. Yoshua Bengio, founder of Element AI, predicted in 2025 that AI decision-making capacities might reach human parity within five years. AI is plainly developing faster than anticipated. Whether AI will evolve into a genuine subject surpassing human capacities remains

DOI: 10.54254/2753-7064/2026.39372

open to debate, yet AI's emergence as an increasingly irreplaceable quasi-subject is already observable. Feasible AI-value-alignment presupposes human readiness to countenance the possibility that future super-intelligent systems may develop independent value judgements and behavioural decisions. Anthropocentrism should give way to mutual-respect-based consensus-building, and a dual-subject human-AI structure should establish the foundational stance for value alignment. Equal-status interaction renders reverse-value-alignment phenomena visible, creating governance opportunities for human agents. Acknowledging AI's quasi-subject status remains nonetheless a conceptual-theoretical stance, not permission for unbounded AI expansion. Recognition of AI's quasi-subject position serves instrumentally to refine value-alignment paradigms and safeguard human existential security. Accordingly, alongside constructing platforms for equal human-AI dialogue, intensified monitoring, constraint, sanction and disciplinary mechanisms are indispensable. AI ought to be treated as quasi-social actors subject to joint oversight by humans and other AI systems, closing responsibility-vacuum zones within human-AI coexistence.

3.2. Embodied-cognitive alignment

In Treatise on Sensations published in 1754, Étienne Bonnot de Condillac described a statue successively endowed with sight, hearing, touch, smell and taste. Acquiring sensory faculties enables the statue to develop perceptual capacities and cultivate attention, memory, comparison, judgement, imagination, recognition and even desire. David Chalmers likewise argues: "If dynamic cognition is computable, then appropriate computational organisation gives rise to consciousness." [12]. Embodiment accordingly constitutes the foundation for artificial-intelligence systems to progress from digital computation toward experiential perception. Most present-day AI lacks embodied-interaction capabilities, creating inherent asymmetries within human-AI communication. Values known to AI consist merely of flat symbolic inputs from humans. Symbols may trigger specific behavioural outputs but cannot replicate human value-acquisition pathways or generate experiential insight for AI, thereby trapping value-alignment efforts within persistent deadlock. Embodied intelligence was incorporated into China's 2025 Government Work Report, which listed artificial intelligence alongside bio-manufacturing and quantum science as priority future-industry sectors. Policy plans envisage integrating AI within physical-robot platforms to endow systems with human-like perception, interaction and learning faculties. Essentially, enhanced agent-environment interaction augments AI initiative and adaptability within physical-world contexts. Anthropomorphic physical architectures confer embodiment, modularity and emergent-property characteristics upon AI. Deep immersion within physical-environments, human-societal and economic-system processes enables AI directly to participate in social-production-and-reproduction activities and develop distinctive social attributes [13], facilitating richer comprehension of human values. When human-AI communication expands beyond textual outputs to incorporate bodily gestures, vocal-emotional cues and simulated facial expressions, genuine value alignment becomes practically conceivable.

3.3. Full-lifecycle alignment

Throughout human-value history, human values have been shaped by humanity's awareness of life's transience and fragility. Human finitude determines practical-interest orientations and risk judgements, giving rise to value-systems oriented toward welfare-maximisation. The AI lifecycle encompasses requirement definition, computing-infrastructure deployment, end-user-device diffusion, monitoring-optimisation and decommissioning — commonly referred to as the full AI lifecycle. Human lifespans are measured chronologically; AI lifecycles are evaluated via social-

DOI: 10.54254/2753-7064/2026.39372

influence metrics, computational-resource consumption and processing-capacity benchmarks. Such lifecycle-incommensurability creates formidable barriers for human-AI value alignment. Even though iteratively updated AI models often possess shorter practical service-lives than human beings, their implicit "infinite-life" design premises prevent AI from adopting human-style value-orientations. Value-alignment implementation consequently requires lifecycle-oriented constraints upon AI. In 2026, Fang Anran and Chaolemen observed in their research on generative-AI self-training loops that generative-artificial-intelligence performance continuously degrades during self-training cycles: generated-content diversity diminishes and human-AI textual-disparities widen [14]. This degradation stems from over-fitting within generative-AI systems when each training iteration exclusively consumes outputs produced by prior-generation models. Insufficient injection of real-world external data narrows expressive scope and erodes linguistic creativity. This research offers actionable insights for designing finite-lifespan AI systems. Given that AI iteration depends upon expanding algorithmic infrastructures and large-scale data-feeding, this paper proposes two design principles for AI lifecycles. First, regulate external-data-access frequencies at distinct lifecycle stages so that AI updating velocity, computational-resource consumption and creative-capacity evolve along parabolic trajectories resembling human-age-related cognitive-capacity shifts. Second, introduce targeted-removal mechanisms that impose resource-access time-limits for AI. Borrowed from medical terminology, "targeted removal" originally describes precise-recognition-based clearance of specific biological substances. Within this paper's framework, it denotes performance-regulation mechanisms that gradually curtail AI's external-data-access privileges over time. AI service-durations are thereby prevented from extending indefinitely; systems undergo regular degradation and eventual deactivation, simulating finite human-like lifecycles. This not only emulates human-biological characteristics to realise structural lifecycle alignment, but also imposes institutional-framework constraints mitigating potential AI-malfunction risks.

4. Conclusion

Artificial-intelligence value alignment is not merely an algorithmic challenge but a philosophical issue concerning social order and humanity's future. It represents both an engineering endeavour and a civilisational project. This paper advocates subject-identity alignment, embodied-cognitive alignment and full-lifecycle alignment between humans and AI. These proposals do not constitute technological veneration of fast-advancing artificial intelligence within a new-technological-revolution context, nor do they amount to anthropocentric bodily-discipline or life-restriction measures imposed upon AI. Instead, they constitute a realistic response to contemporary technological trajectories. Through achieving human-AI value alignment, we seek to build shared human-AI value-communities and realise symbiotic, coexistent and integrated human-AI relationships. Investigating artificial-intelligence value alignment amounts simultaneously to value-enlightenment for AI and normative-civilisational self-reflection for humanity. On pathways out of value-alignment predicaments, alignment is required not only for technologies themselves, but also for human understandings of our own civilisational heritage.

References

[1] Betley, J., Tan, D., Warncke, N., et al. (2025). Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. https://icml.cc/virtual/2025/poster/44803

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

DOI: 10.54254/2753-7064/2026.39372

[3] Lin, A., & Chang, Y. (2024). Humanistic reflections on value alignment of large artificial intelligence models. *Journalism Review, 8*, 24–33.
[4] Sun, K., & Wang, J. (2026). Reconstructing artificial intelligence governance paradigms under the reverse value alignment dilemma. *Journal of Dialectics of Nature, 48*(5), 101–109.
[5] Hume, D. (2016). *A treatise of human nature*. (W. Guan, Trans.). The Commercial Press.
[6] Marx, K., & Engels, F. (2009). *Collected works of Marx and Engels (Vol. 8)*. People's Publishing House.
[7] Heidegger, M. (2018). *Lectures and essays* (Z. Sun, Trans.). The Commercial Press.
[8] Yao, X. (2026). An analysis of the dissolution and defence of human subjectivity from the perspective of artificial intelligence instrumentalism. *Theory Journal, 3*, 76–85.
[9] Xue, G., & Zhao, J. (2023). From 'ought' to 'is': Exploring practical strategies for artificial intelligence ethical norm implementation. *Studies in Dialectics of Nature, 39*(1), 88–96.
[10] Yan, H., & Li, Y. (2024). Exploring coping logics for deceptive value alignment: From 'intention' to 'symbiosis'. *Journal of Huazhong University of Science and Technology (Social Science Edition), 38*(5), 20–28.
[11] Kurzweil, R. (2011). *The singularity is near*. Li, Q. (trans.). China Machine Press.
[12] Chalmers, D. J. (2012). *The conscious mind*. Zhu, J. (trans.). China Renmin University Press.
[13] Jiang, H. (2025). "When statues can perceive": An inquiry into the possibility of artificial intelligence consciousness. *Nanjing Social Sciences*, (8), 117–128.
[14] Fang, A. R., & Chao, L. M. Model degradation risk in self-training of generative artificial intelligence. *Computer Engineering*, 1–9(Advance online publication; accessed 2026-09-22).