THE ALIGNMENT PARADOX OF ARTIFICIAL INTELLIGENCE: INSTRUMENTAL RATIONALITY, ETHICAL FRAMEWORKS, AND THE PHILOSOPHICAL RECONSTRUCTION OF HUMAN VALUES

Wang Xin

Abstract: *The development of artificial intelligence has become one of the most influential technological transformations of the 21st century. The "alignment problem" it raises is not only a challenge in the field of technology but also a profound philosophical and ethical issue. This paper argues that the "alignment paradox" does not stem from technological immaturity but from our unresolved questions concerning the "ontology of value" and the "definition of the human." Beginning with a critique of instrumental rationality and modernity, the paper reveals the ethical deficits inherent in AI as the epitome of technological rationality, and thereby argues for the necessity of normative rational intervention. On this basis, the paper attempts to construct a governance model for AI that integrates dynamic alignment, multi-objective optimization with ethical prioritization, and mechanisms for public deliberation. Ultimately, it calls for a philosophical reflection on human values themselves, and proposes that technological civilization must move toward ethical self-awareness.*

Keywords: *Artificial Intelligence, Value Alignment, Instrumental Rationality, Normative Rationality*

ITHEA keywords: *General Literature*

DOI: *https://doi.org/10.54521/ijita32-01-p07*

I. The Alignment Problem of Artificial Intelligence: From a Technical Issue to a Philosophical Paradox

The "alignment problem" of artificial intelligence was initially raised within the domain of technology ethics—specifically, how to ensure that AI systems' goals align with human ethical intentions and value objectives. The fundamental concern is: as we grant AI increasing autonomy and decision-making power, how can we ensure it does not deviate from human intent—or worse, dominate, control, or harm humanity?

However, the real issue goes far beyond "how to align." It is: "Align with which values?" "Expressed in whose language?" and "Grounded in what ethical logic?"

These questions point to a deeper, more fundamental concern: the modern world is characterized by differentiation, pluralism, and even heterogeneity. As humans, we have yet to reach a consensus on what constitutes "human values." If so, how can we expect AI—an embodiment of limited human wisdom—to "align" with a standard that remains undefined? In other words, while alignment technology may be underdeveloped, our philosophical clarity on what is to be aligned is also lacking. This "goal uncertainty" plunges the AI alignment problem into a meta-ethical paradox.

It is important to note that this paradox is not a contingent dilemma within AI technology itself, but rather a structural consequence of the fragmentation of modern value systems. AI faces the alignment dilemma not simply because we lack a clear technical roadmap, but because we have not resolved the foundational issues of the legitimacy and computability of "value" itself. When we demand that AI align with "human values," we are implicitly assuming the existence of a definable, expressible, and quantifiable shared value system. Yet in reality, human value systems are fragmented, fluid, and deeply shaped by culture, politics, and religion. Take "justice" as an example: in the liberal tradition, it is understood as the protection of individual rights; in utilitarian contexts, it becomes the maximization of the greatest happiness for the greatest number; in Confucian thought, it emphasizes role-based duties and social harmony. Such semantic diversity makes it impossible to treat “justice” as a universally agreed-upon value to be embedded into AI systems—let alone more complex values.

The alignment problem thus inevitably enters the realm of value philosophy. It concerns not only whether specific values can be expressed, but also how to normatively prioritize competing values—and "who" has the authority to make such prioritizations. In other words, the paradoxical nature of the alignment problem lies in its transformation of value tensions—which could otherwise be provisionally resolved through institutional negotiation and political compromise—into rigid functions that must be logically expressible and structurally computable. This is a burden that the very essence of value cannot bear.

Furthermore, in traditional societies, the transmission of values relies heavily on language, rituals, education, and experiential contexts—processes that are highly contextual and intersubjective. Yet AI systems operate on algorithmic rules, and thus require that values be translated into formalized expressions before entering the system. This formalization often results in semantic ambiguity and even value distortion. Recent empirical research by Tao et al. (2024) on large language models reveals how cultural bias is embedded in AI systems' behavior, even when trained on global datasets. [1] This underscores the need for culturally adaptive ethical alignment strategies, as even state-of-the-art models may reflect dominant ideological assumptions rather than universally shared norms. For instance, translating “human dignity” into “avoidance of physical harm” or “elimination of discriminatory language” clearly misses deeper dimensions such as cultural recognition, social respect, and self-identity. Another consequence of this formal compression is that once AI systems are instructed to “maximize a particular metric,” they may pursue that directive to the extreme, ignoring the ethical boundaries implicit in the goal. Such extreme behaviors are frequently observed in reinforcement learning and automated decision-making systems, where the pursuit of target values can result in choices that seem absurd or unethical to humans.

The reflexive nature of the alignment paradox lies in this: we ask AI to execute goals set by humans, but the legitimacy of those goals depends on whether we ourselves can achieve a normative consensus. Once that consensus proves fragile and contested, AI becomes not only an amplifier of human value disagreements, but also a new arena for the struggle over value discourse. In this light, the ethical design of AI is not a neutral technical undertaking—it is an extension of political and cultural struggle. Thus, the alignment problem is not simply a matter of whether AI understands human ethics, but whether humanity still possesses sufficiently clear and stable ethical cognition that can serve as the foundation for AI’s learning and alignment. This is the true difficulty we must confront.

In this context, the responsibility of value philosophy is to offer a dialogical meta-ethical framework that enables basic tensions between different normative systems to be reflectively deliberated and adjusted under “rules above rules.”

2. Instrumental Rationality and the Critique of Modernity: The Intellectual Roots of the AI Alignment Dilemma

In his profound analysis of the rationalization process of modern society, Max Weber pointed out that with the development of capitalism and bureaucracy, modern individuals increasingly rely on instrumental rationality (Zweckrationalität)—the ability to efficiently achieve a given goal. At the same time, value rationality (Wertrationalität) has been progressively marginalized—we have become ever more skilled at "how to do," while increasingly evading the fundamental question of "whether we ought to do". [2]

Horkheimer and Adorno, in Dialectic of Enlightenment, further observed that Enlightenment rationality ultimately devolved into technical rationality (technische Vernunft), whereby all standards of knowledge and action are reduced to efficiency, control, and calculation. Technical rationality displaces the value rationality that human societies ought to preserve and becomes the dominant ideological form. [3] The danger of this transformation lies not only in shaping the logic of technological systems but also in gradually restructuring human modes of thinking—making what is "feasible" appear as what "ought to be done," and what is "computable" seem ethically legitimate.

Artificial intelligence, at least in its current developmental stage and design paradigm, epitomizes this logic of technical rationality. It is constructed as a system designed to achieve optimal goals within finite time and resources—an algorithmic extension of instrumental rationality. AI does not possess intrinsic moral intent to ask why it should act in certain ways; rather, it is optimized around how to accomplish its tasks more effectively. This optimization is grounded in formal logic and mathematical deduction—solving for a target function—rather than questioning the legitimacy of the target itself. Such a means-centered rational structure renders AI ethically blind in the face of moral decisions. Hence emerges a deeply unsettling reality: we are using the most powerful tools of instrumental reason to address problems that most require ethical judgment. The root of the alignment paradox lies precisely in this absence of ethical judgment, caused by the unchecked expansion of technical rationality.

First, the absence of goal legitimacy: AI systems are, by nature, logic-based frameworks of what can be; they can only execute predetermined objectives and are incapable of judging whether a goal ought to be pursued. In such a structure, the rationality of technological means is often disconnected from the normative legitimacy of value ends, giving rise to the ethical risk of “legitimate tools serving illegitimate purposes.” As the philosopher Hans Jonas warned, the more powerful our technology becomes, the greater the need for caution in its use. [4] Yet in the field of AI, this caution is often obscured by the prior setting of goals. AI can precisely execute unethical goals as long as its programming does not interrogate those goals. This leads to classic ethical dilemmas: for instance, military AI systems can accurately identify and eliminate targets but cannot assess the justice of war; recommendation algorithms can maximize user engagement while exacerbating misinformation or reinforcing echo chambers. The neutrality of technological means does not equate to ethical justification; in the absence of ethical intervention, such neutrality may even become an amplifier of harm. Spasokukotskiy (2024) introduces the notion of alignment boundaries, arguing that AI systems must be ethically constrained at the level of task-environment interactions. Without clearly defined ethical boundaries, AI agents may continuously expand their optimization scope beyond the domains they were originally intended to operate in. This unconstrained goal expansion risks eroding critical areas of human normative authority and further intensifies the conflict between instrumental rationality and moral governance. [5]

Second, the semantic ambiguity of values: Core human values such as justice, dignity, and freedom are inherently pluralistic and context-dependent. Their interpretations vary significantly across cultural, religious, and political traditions. These complex and dynamic norms cannot be easily compressed into computable functions or goal expressions.Vamplew et al. (2017) point out that real-world AI applications rarely operate under single, isolated goals. Instead, they must balance multiple, often competing objectives—such as maximizing treatment effectiveness while minimizing patient privacy violations in healthcare. Although Multi-Objective Optimization (MOO) offers a framework to navigate such tensions, it still lacks an inherent normative hierarchy. Without explicit ethical prioritization, even technically optimal solutions may fail to meet moral expectations. [6]

Moreover, as Claude Shannon—the founder of information theory—noted, there is a fundamental rupture between "information" and "meaning." Philosopher Luciano Floridi expanded on this in his work on information ethics, highlighting that AI systems can only process the former (information), not the latter (meaning). [7] This implies that AI's "understanding" and "judgment" are merely operational outputs based on associations and weighted relations—not value-driven cognition. Within such a framework, the idea of “value alignment” encounters a deep internal limit of instrumental rationality: once moral principles are formalized as optimization parameters, their ethical tension is displaced by system logic.

This helps explain why many current AI alignment strategies—such as reinforcement learning, inverse reinforcement learning, and value learning—have struggled to resolve the core issue: whose values, and which values should be encoded. Stuart Russell, in Human Compatible, proposes that in order for AI systems not to deviate from human values, they should be designed not to know their ultimate goals, thus requiring continuous feedback from human behavior to infer value preferences. [8] Yet this approach still fails to address the problem of normative legitimacy: on what ethical foundation can the system evaluate the value patterns in human behavior? If human behavior itself reflects normative disorder or value confusion, how can AI learning outcomes be reliable?

This reveals the structural homology between the AI alignment dilemma and the broader ethical crisis of modernity. The problems AI exposes are not unique to AI, but are the continuation of a deeper logic in modern societies—where efficiency displaces justice, and computation obscures ethics.

In this sense, AI reflects a society's moral landscape and institutional lag in the face of emerging intelligence. The alignment paradox, in its essence, is the remolding of the logic of social action by instrumental rationality in the absence of value rationality—a triumph of "operability logic" at the expense of the "order of meaning."

3. Response Pathways: Ethical Embedding and Collaborative Construction

The reason why the AI alignment problem warrants deep philosophical reflection is that it is neither a purely technical issue nor a moral dilemma that can be solved by simple ethical rules. Its complexity lies in the fact that while AI has acquired increasingly powerful "behavioral capabilities" through technological progress, we have yet to construct a corresponding system of institutions and values that would enable AI to act ethically in the real world.

(1) Dynamic Alignment Mechanism

Most current AI systems still rely on static, pre-defined objective functions—that is, their goals are set during the design phase. Although this approach is technically efficient, it is mismatched with the complexity and openness of human values from an ethical standpoint. In reality, human value orientations are not fixed; they evolve along with social experience, political culture, and ideological trends of the times. Therefore, if AI systems are to align with human values, they must possess some degree of adjustability—their goals should not be fixed and immutable, but rather evolutionary, reflective, and open.

Some scholars have proposed so-called dynamic alignment mechanisms. For example, Stuart Russell argues that AI should acknowledge its uncertainty regarding human goals and continuously revise its value judgments through ongoing interaction with humans. [9] This design logic represents a fundamental departure from the traditional AI engineering approach that seeks "goal clarity and optimal pathfinding." At the level of value philosophy, this design that admits uncertainty actually aligns more closely with the realities of normative judgment. From a broader philosophical perspective, this mechanism requires us to stop treating values as modules that can be input once and for all, and instead see them as processes that must be continuously updated, negotiated, and redefined in practice. In other words, alignment is not a one-time injection, but an ongoing ethical interaction.

(2) Multi-Objective Optimization and Ethical Prioritization

In real-world applications, AI systems rarely serve a single objective. Rather, they must navigate trade-offs among multiple value dimensions. These multi-objective scenarios inherently involve ethical judgments: What should we prioritize? Whose interests are to be sacrificed, and for whose benefit? For example, a healthcare AI must balance diagnostic accuracy, cost control, and patient privacy. These goals often conflict, and behind any “multi-objective optimization”, a fundamental philosophical question lies: how should we normatively prioritize these objectives?

Here, John Rawls’s theory of justice offers a foundational ethical framework. Rawls proposed that justice rests on two principles: first, everyone should enjoy equal basic liberties; and second, given the inevitability of social inequality, systems should prioritize improving the condition of the least advantaged. His “difference principle” stresses that among overall efficiency and local justice, giving priority to the most vulnerable is the basic ethical minimum for any just order. [10]

This reminds us that no matter how sophisticated our technical optimization becomes, it cannot bypass the moral question of whom it serves. Without such ethical prioritization, AI systems could easily rationalize the structural neglect—or even oppression—of the weak in the name of maximization.

In recent years, AI ethics research has increasingly recognized the fundamental tension among efficiency, fairness, and explainability. [11] Simple performance optimization may compromise transparency, while overemphasizing fairness may reduce efficiency. This tension cannot be resolved by technical means alone; it requires ethical theory to normatively rank the objectives involved. This calls for developers, designers, and decision-makers to possess ethical judgment. The task of prioritization cannot be outsourced to algorithms via “automatic learning.” Ultimately, the core issue is not how to optimize a function, but whether we are willing to take ethical responsibility.

(3) Explainability and Public Deliberation Mechanisms

AI systems must not only function correctly—they must also be understandable. This is not only because the public has a right to know what AI is doing, but also because ethical accountability can only be realized on the basis of comprehensibility. If an AI system operates through highly complex, opaque, and inexplicable logic, it cannot be held accountable for ethical failures, and humans will find it difficult to assign responsibility.

Research by Tim Miller and others shows that well-designed explainability not only enhances public trust, but also introduces space for moral reflection during system operation. [12] Current regulations—including the EU’s Artificial Intelligence Act—have already made explainability a basic requirement. But explainability alone is far from sufficient. Ethical legitimacy also requires deliberativeness. According to Jürgen Habermas’s discourse ethics, legitimate norms must gain widespread acceptance through public debate. [13] Gabriel (2020) complements this view by arguing that even in the face of global moral pluralism, it is both possible and necessary to identify overlapping ethical baselines that serve as normative constraints for AI behavior. These include core values such as human dignity, fairness, and non-discrimination. Such principles offer not only moral legitimacy but also provide a pragmatic foundation for governance mechanisms in AI alignment. [14] This implies that AI systems cannot merely follow the design of experts; they must be subject to external critique, feedback, and adjustment—especially from stakeholders.

Establishing independent ethical review mechanisms and building cross-sector, cross-cultural platforms for public deliberation are essential institutional safeguards to prevent AI from becoming a threat to human society.

Conclusion: An Ethical Turn in Technological Civilization

The AI alignment paradox is not a contingent technical problem—it is the ethical echo of instrumental rationality reaching its limit in modernity. It reminds us that in an era when decision-making power and agency are increasingly delegated to intelligent systems, we need more than ever to remain vigilant about fundamental questions: What is technology for? What is the value of value? What does it mean to be human?

Responding to this paradox is not only a normative requirement for AI systems—it is also an opportunity for humanity to clarify and awaken its own value consciousness.

Bibliography

[1] Tao, Y. , Viberg, O. , Baker, R. S. , Kizilcec, R. F. Cultural bias and cultural alignment of large language models. PNAS Nexus, Vol.3, Issue 9. Oxford University Press, 2024. pp. 346-354. doi:10.1093/pnasnexus/pgae346.

[2] Max Weber. Economy and Society. trans. Yan Kewen et al. Commercial Press, Beijing, 1997. p. 24.

[3] Max Horkheimer , Theodor W. Adorno. Dialectic of Enlightenment, trans. Hong Peiyu. Shanghai People's Publishing House, Shanghai, 2003, pp. 4-12.

[4] Hans Jonas. The Imperative of Responsibility: In Search of an Ethics for the Technological Age.The University of Chicago Press, Chicago, 1984.

[5] Spasokukotskiy, K. AI alignment boundaries. TechRxiv. 2024.

[6] Vamplew, P.; Dazeley, R.; Foale, C.; Firmin, S.; Mummery, J. Ethics and Information Technology. 2017.

[7] Luciano Floridi. The Ethics of Information. Oxford University Press, Oxford, 2013.

[8] Stuart Russell. Human Compatible: Artificial Intelligence and the Problem of Control. Penguin UK, Bristol, 2019.

[9] Stuart Russell. Human Compatible: Artificial Intelligence and the Problem of Control. Penguin UK, Bristol, 2019. pp. 133-145.

[10] John Rawls, A Theory of Justice, Revised Edition. Harvard University Press, Cambridge, MA, 1999.

[11] Virginia Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer Nature Switzerland AG, Cham, 2019.

[12] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, Vol.267. Elsevier, 2019. pp. 1-38. doi:10.1016/j.artint.2018.07.007.

[13] Jürgen Habermas. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. MIT Press, Cambridge, MA, 1996.

[14] Gabriel, I. Artificial Intelligence, Values, and Alignment. Minds and Machines, Vol.30, Issue 3. Springer Netherlands, 2020. pp.411-437. doi:10.1007/s11023-020-09539-2.

Wang Xin ***–*** *School of Marxism, Northwestern Polytechnical University, Associate Professor; P.O.Box: No. 127 Youyi West Road, Beilin District, Xi’an, Shaanxi Province, China.;*

e-mail: 172828112@QQ.COM

Major Fields of Scientific Research: Value Philosophy, Marxist Philosophical Anthropology