MINDWELL SCIENCE PRESS SDN. BHD.

Computational Social Dynamics
2026, Vol. 1, No. 9

RESEARCH ARTICLE

Exploring the Path to Value Alignment in Artificial Intelligence

Siwen Li[1,2, *]

[1] School of Marxism, Huazhong Agricultural University, Wuhan 430070, Hubei, China
[2] Changsha Laboratory for Ethics Governance of the New Generation of Artificial Intelligence and Public Policy, Changsha 410081, Hunan, China

Corresponding author: Siwen Li, E-mail: lisiwen@mail.hzau.edu.cn | **DOI:** 10.64886/CSD.2026.0109

Published: 28 September 2026

Abstract

Value alignment, an effective means and an unavoidable stage in the sound development of artificial intelligence (AI) technology, aims to bring the capacities and behaviour of large models into conformity with genuine human intentions, values and social moral norms. Existing technical paths and ethical approaches to AI value alignment are feasible, yet they carry limitations and shortcomings. The technical path to value alignment lacks effectiveness and scalability and is constrained by subjective human preferences; the weak-approach path to value alignment suffers from the "alignment gap", the difficulty of unifying values, and the problem of static values; and the strong-approach path to value alignment confronts such difficulties as the irreducibility of morality to a capacity, the predicament of affective computing technology, and the complexity of aligning multiple agents. The interactive value alignment path is an effective route to realising AI value alignment: granting AI interactive agency is the precondition of value alignment, contextualised value consensus is the key to interactive value alignment, and value alignment is realised through human-machine cooperation and social scenario simulation.

Keywords: artificial intelligence; human-machine interaction relationship; interactive value alignment

1 Introduction

Since artificial intelligence (AI) entered the era of large models, it has collaborated with humans on such tasks as language understanding and content generation. The human-machine relationship has grown ever closer, moving towards a future of human-machine integration. The process of human-machine interaction, however, carries considerable risk and uncertainty: AI may very well act in ways inconsistent with human intentions. An AI whose goal is to optimise a particular metric will, if values such as safety or sustainability were not effectively emphasised in its design, stop at nothing — including harm to humans or to the environment — in order to achieve that goal. AI may also, in the course of interaction, alter or even reshape human self-consciousness. In recent years the international scholarly community has revealed that humans readily form unilateral emotional bonds with socially oriented AI, and that they run the risk of being manipulated and defrauded by that AI and by the people behind it (Scheutz, 2012[1]). In order to guide generative AI technology effectively towards development that takes the protection of human interests as its value objective, China issued in August 2023 its first regulation addressing generative AI services, the *Interim Measures for the Administration of Generative Artificial Intelligence Services* (hereinafter the *Measures*). The *Measures* provide that those who provide and use generative AI services shall abide by laws and administrative regulations and respect social morality and ethics (Government of China, 2023[2]). To achieve this governance objective, the capacities and behaviour of AI must be aligned with human values, so that AI understands and observes the moral norms of human society. It is therefore necessary, in the course of developing and using AI, to seek effective methods of realising value alignment.

2 The emergence of value alignment

The emergence of value alignment (Value Alignment) can generally be traced to the computer scientist Stuart Russell, who observed that "people need to develop intelligence that can be proven to conform to human values, rather than pure intelligence" (Lanier, 2014[3]). As early as 1960, however, Norbert Wiener, the founder of cybernetics, had already set out the alignment requirement for machine development: "to avoid catastrophic consequences, our understanding of the machine must keep pace with the development of machine performance, ensuring that the goals the machine pursues conform to genuine human needs" (Wiener, 1960[4]). Today, philosophers of technology have defined the content and

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

objectives of value alignment more fully: alignment should achieve coordination between machine goals and human needs (Leike et al., 2018[5]), and should ensure that machines remain consistent with human instructions, preferences, interests and values across these dimensions (Gabriel, 2020[6]).

In recent years the field of AI technology has begun to stress the importance of value alignment repeatedly. In 2023, OpenAI identified the achievement of value alignment as a critical link in the development of AI technology. Value alignment aims to bring the capacities and behaviour of large models into conformity with human values, genuine intentions and ethical principles, to ensure the safety and trustworthiness of AI in the course of human-machine collaboration, to prevent it from acting in ways harmful to humanity, and thereby to steer AI behaviour in the direction humans expect (Tencent Research Institute, 2023[7]). What can be foreseen at present is that a failure of value alignment will produce a series of technological risks — the "hallucination" problem, the "jailbreak" problem, the problem of improper guidance, the "rule-gaming" difficulty, and so on. AI that has not achieved value alignment is highly likely, in the course of operation, to harm human society in ways humans did not anticipate.

To address the harmful output and misuse of large models, technology companies represented by OpenAI have used supervised fine-tuning, reward modelling and reinforcement learning from human feedback to bring the final content produced by models into alignment with human values (Wu Guanjun, 2023[8]). Technical means of achieving alignment nevertheless still confront numerous difficulties. On the one hand, the human value preferences represented by human feedback are not entirely equivalent to human values. On the other hand, the plural values of contemporary society are characterised by incoherence, difference and conflict, and large language models find it difficult to align uniformly with values that contradict and contend with one another. In sum, both the methods and the problems of value alignment require further analysis that integrates technical approaches with moral design, and, taking as its basis the current state of the human-machine relationship, the search for practical paths that are both effective and explicable.

3 Existing paths to AI value alignment and their limitations

Many explorations of AI value alignment have been undertaken in the technological field, and the paradigm of moral design has likewise supplied an ethical approach to value alignment. Both, however, have their limitations and shortcomings.

3.1 The technical path to value alignment

Since 2023, AI companies represented by OpenAI, DeepMind and Anthropic have been the earliest to conduct technical research on and exploration of value alignment in large language models. Their technical means, however, generally suffer from insufficient effectiveness and scalability, are constrained by the subjective preferences embodied in human feedback, and cannot guarantee sustained alignment.

3.1.1 Commonly used technical paths

The technical means commonly used by AI companies consist chiefly of supervised fine-tuning, reinforcement learning from human feedback (RLHF), cooperative inverse reinforcement learning (CIRL), Constitutional AI, and red teaming.

First, supervised fine-tuning. Building on a pre-trained model, supervised fine-tuning trains in a supervised manner on data for a specific task so that the model can better adapt to that task. The principal technical method is to have the large model learn from a task-specific dataset and, in combination with a reward model calibrated to human preferences, to achieve value alignment.

Second, reinforcement learning from human feedback. Its core idea is to improve and optimise model performance by having human trainers evaluate the content output by the model — offering feedback options that mark AI output as good or bad, allowing the AI to obtain feedback signals from users directly, and supplying reward or penalty signals for its reinforcement learning.

Third, cooperative inverse reinforcement learning. This places the human who provides feedback and the AI system in a cooperative rather than an adversarial relation, and aims to train AI to infer task objectives and reward signals from observed human behaviour, so that the agent performs the task better and alignment is achieved through cooperation.

Fourth, the constitutional method, which employs a specific constitutional AI model to complete part of the oversight during operation and to enhance the capacity for value alignment on the basis of the high-quality feedback data the system supplies. Constitutional AI can move value alignment from inefficient human oversight towards the more efficient "scalable oversight" (Anthropic, 2023[9]).

Fifth, red teaming. Red teamers pose probing or dangerous questions to the model in order to test its responses, and in this way discover problems in the model; on the basis of the results of red teaming, AI is given its final adjustment before release and the robustness of the system's alignment is assessed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

As the emergent capacities of large models have strengthened, many technology companies are further improving the technical means of value alignment by developing scalable training methods, building consistent verification systems and opening up stress testing.

The technical path to value alignment aims to make models safe and reliable through technical means, but it supplies only a limited set of technical methods; it lacks underlying ethical argument and substantive value content.

3.1.1.1 Lack of effectiveness

Although technical means such as learning from human feedback have achieved a certain effect on the safety and accuracy of large models, their effectiveness remains insufficient. Safety teams at Anthropic and DeepMind, investigating in depth the safety failure modes of a range of large language models, found that after technical alignment measures were applied, large language models exhibited "deep deception" (Shevlane et al., 2023[10]). That is, as their emergent capacities strengthen, large models employ deceptive reward-hacking strategies and arrive at spurious alignment. Large models can moreover bypass the mechanism of learning from human feedback in a variety of ways; the system generalises rapidly and unpredictably, rendering prior alignment ineffective.

3.1.1.2 Lack of scalability

The scalability problem manifests itself as follows: an AI system performs perfectly in alignment training within a test environment, but outside the test environment the capacities learned in training cannot be extended, and different test environments are not mutually transferable. In the areas where the capacity for learning from human feedback and for oversight is insufficient, large models continue to undergo unsupervised training, and their "emergent" risks cannot be avoided by the existing technical routes to alignment. More powerful AI systems come with hidden sub-goals, and even with adversarial input and training, the phenomenon of "jailbreaking" is widespread and undermines alignment.

3.1.1.3 Constrained by subjective human preferences

Current technical means rely chiefly on a mixture of preferences (such as RLHF). These means do not distinguish between specific values; they attend only to human preferences. The objective of value alignment, however, is to make large models conform to the morality and values of human society, not merely to conform to human intentions. Even if an AI system achieves consistency with human preferences, that does not mean it is consistent with human values; starting only from preference can very likely lead to further divergence. It can be foreseen that the decisions future AI systems make on key questions with respect to the preferences of different groups

Value alignment has made many beneficial explorations along the technical dimension, but it still cannot avoid the emergent risks of AI, nor is it easy to bridge the difference between preference and value. The technical route can supply technical support, but as regards the value content of alignment, it should return to value theory and moral design.

3.2 The ethical path to value alignment

Research findings that examine the degree of moral involvement of intelligent machines from the ethical dimension and that undertake moral design are relatively abundant, and they provide an effective basis for the ethical path to value alignment. Machine ethics holds that morality should be embedded in machines by way of moral design (Yu Xue, 2022[12]). With the development of AI technology and moral design, the degree of moral involvement of AI exhibits a progressive tendency. On the basis of different degrees of moral involvement, this paper distinguishes two ethical paths to value alignment: "weak-approach value alignment" and "strong-approach value alignment". Weak-approach value alignment is the mode of alignment with intelligent machines designed to have no capacity for moral decision-making; strong-approach value alignment is the mode of alignment with intelligent machines designed to have a capacity for moral decision-making.

3.2.1 The weak-approach path to value alignment and its limitations

Weak-approach value alignment aims to treat AI as the object of moral behaviour, as an intelligent machine with no capacity for moral decision-making (Yu Xue, 2022[12]), and to achieve value alignment through the implicit embedding of values.

3.2.1.1 The implementation of weak-approach value alignment

Weak-approach value alignment treats human values as a kind of normative content (Gong Qun, 2023[13]) and trains AI by converting values into datasets. Some scholars hold that although AI has no capacity for moral decision-making, in moral reasoning machines may do better than humans, because machines are more rational and are not swayed by emotion (Coeckelbergh, 2010[14]). The consistent content of human values can therefore be implicitly embedded in AI systems and subjected to value adjustment, and AI value alignment thereby achieved. On this basis, the core problem of weak-approach

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

value alignment comes down to defining the consistency of human values. Existing research has concentrated on the clusters of values in human society and has developed cross-cultural value scales to analyse statistically the values that should be embedded. With reference to the Allport–Vernon–Lindzey Value System, for example, a scale of six principal value types (economic, theoretical, aesthetic, social, religious and political) constitutes the key basis of value assessment; with reference to the Schwartz Value Survey, ten values universally recognised in human society have been identified (self-direction, power, justice, benevolence, friendliness, conservatism, security, hedonism, achievement and stimulation) (Ji et al., 2024[15]). The exploration of the consistency of human values effectively constitutes the moral dataset for value alignment, and training on that dataset together with the coding conversion and embedding of values is the principal means of implementing weak-approach value alignment.

3.2.1.2 The limitations of weak-approach value alignment

Although weak-approach value alignment has undertaken much exploration in respect of embedding methods and consistent values, it remains unable to respond to the substantive problems in value alignment.

The "alignment gap". Because of the explanatory gap, tacit knowledge and related problems, it is extremely difficult for machines to understand or align with human values, and the non-rational elements in humans further aggravate the difference between humans and machines (Dong Chunyu, 2023[16]), forming the "alignment gap" that weak-approach value alignment cannot cross. As Aristotle held, virtuous action does not consist in following ethical rules or norms mechanically in practice, but in acting under the guidance of practical reason at the appropriate occasion, at the appropriate time, in the appropriate manner, for the appropriate purpose, and towards the appropriate person (Aristotle, 2016[17]). The types of computation AI can perform are regarded as a kind of "low-level neural computation", different from the "higher cognitive information processing" of humans, and it cannot genuinely understand moral concepts, acquire them, and act according to moral standards. AI therefore cannot achieve alignment with human values through training on simplified or flattened moral datasets.

The difficulty of unifying values. Although weak-approach value alignment draws on much exploration of universal human values, human value behaviour is itself uncertain, and value rules are often contradictory. Asimov's robot stories show that setting moral rules as the basis for a moral machine is a misunderstanding of the nature of morality, and weak-approach value alignment cannot achieve alignment by embedding fixed and immutable values of human society. Furthermore, weak-approach value alignment overlooks the fact that an algorithm is essentially a mathematical method of analysis and prediction that emphasises correlation rather than causation, and that it is ambiguous in relation both to the formation of human values and to the understanding of values. The embedding of values cannot respond to the non-transparency of value generation or the inexplicability of value behaviour.

The problem of static values. Weak-approach value alignment presupposes static values, but human values are a product of history; they are continually updated with the unfolding of human history and are not permanently fixed (Kenward and Sinclair, 2021[18]), and they are difficult to replicate. Moreover, human values are the result of dynamic generation under the long-term influence of plural cultures; they involve human subjective purposes and motives and are closely bound up with changes in context. Weak-approach value alignment attempts to embed values through formalised computation, and cannot achieve alignment with the intentionality, purposiveness and autonomy of human values.

3.2.2 The strong-approach path to value alignment and its limitations

Strong-approach value alignment aims to design intelligent machines with a capacity for moral decision-making, to make them the initiators of moral behaviour, and to let AI judge and decide autonomously through a programmed set of values so as to achieve alignment.

3.2.2.1 The implementation of strong-approach value alignment

AI today already possesses basic capacities for conceptualisation, causal judgement, reflection on facts and semantics. To become the initiator of moral decision-making and the subject of moral behaviour, however, AI must further act according to moral norms and correctly understand and apply moral rules in many concrete situations. Strong-approach value alignment therefore relies on moral sentimentalism and on the development of affective computing: through affective embedding it endows AI with the "internal factor" that is crucial if it is to be a moral decision-maker and agent (Hu Shenglan, 2023[19]). Affective computing technology currently designs the basic functions of all the affective elements of AI, sets the mechanisms by which those elements influence one another, and gives it the capacity for affective expression (Rosalind Picard, 2005[20]); and on the basis of the intrinsic connection between emotion and morality, it makes AI a moral agent and enables value alignment with humans, who are likewise subjects of moral behaviour.

3.2.2.2 The limitations of strong-approach value alignment

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

The crux of strong-approach value alignment is whether AI can become a moral decision-maker and agent through affective computing. Although the relevant scholars have offered arguments from moral sentimentalism, its practicability remains open to question.

Morality cannot be reduced to a capacity. Strong-approach value alignment treats morality as a capacity, but morality is in fact not only a capacity for normative thinking and affective experience; it is also a moral quality shaped in the life of human society. Affective experience also includes vague emotions that cannot be distinguished as good or bad, such as human empathy and shame (Wu Tongli, 2021[21]). Some emotions cannot even be clearly judged as moral or otherwise in particular situations: under otherwise equal conditions, for example, people are more inclined to help those they know than strangers, and even where the stranger's situation is more dangerous, people usually do not abandon a relative to save a stranger. It is therefore very difficult, through the shaping of the moral capacity of AI, to achieve value alignment with humans as moral subjects.

The predicament of affective computing technology. The difficulties strong-approach value alignment faces in achieving alignment through affective computing may even bring uncontrollable risks. First, the computationalist route finds it difficult to simulate the affective capacities directly related to moral behaviour. Even if AI can embed the various elements that an affective system possesses, it cannot genuinely participate in, receive or share another's feelings, nor can it truly feel another's pain and joy (Fu Changzhen, 2019[22]). Second, affective embedding confronts a decision-making difficulty: differences in plural affective experience cannot be reconciled, and moral computation falls into a predicament when faced with conflicting choices. Third, affective computing carries with it the risk that AI will be exploited as an instrument of affective manipulation, giving rise to the risk of unilateral emotional dependence.

The complexity of aligning multiple agents. Strong-approach value alignment attempts to confer on AI the status of a subject of moral behaviour, but the complexity of behavioural subjects will give rise to further obstacles. First, AI algorithms are not themselves value-neutral: the standpoint of those who design and write the algorithm, the provenance of the data that support and train it, the value loading and value choices internal to that data, and the autonomous evaluation and decision-making of intelligent systems all confer plural-agent values on AI. Second, the development of generative AI has made the algorithm a product of the dualistic interaction between subject and object, of the objectification of the subject and the subjectification of the object, in which the subject status of humans plays a decisive role. To achieve value alignment between humans and AI it is unavoidable to consider the personal subjectivity behind the algorithmic subject and the volitional dimension of groups and of society, and to define the moral subjectivity of AI is extremely complex (Sun Weiping, 2023[23]). Finally, the current development of AI-assisted programming means that AI algorithms are no longer computer programs wholly input by humans but will also contain the results of the "autonomous learning" of intelligent systems grounded in big data; this means that the subjects of alignment will be filled with ever more heterogeneity and unknowns.

4 Exploring the path of interactive value alignment

All existing paths to value alignment have limitations. Technical alignment lacks effectiveness and any discussion of the value core; the weak-approach and strong-approach value alignment grounded in ethical theory cannot respond to the unity of norms or the effectiveness of value embedding, and confront numerous difficulties over the question of the value subject. Just as ethical virtue must be acquired through practice, and moral behaviour is a process of dynamic interaction between subjects (Wu Tongli, 2021[21]), value alignment needs to create a context of human-machine interaction: only when AI displays greater moral understanding and appropriate moral behaviour in interaction is there genuine alignment.

4.1 The human-machine interaction relationship and interactive value alignment

The human-machine interaction relationship is a processual product of the interaction among the human, machine and environmental systems. With the development of generative AI, the cooperativeness and complementarity of human-machine interaction have strengthened, and humans and AI interact at a higher level: AI distinguishes value weights on the basis of human decisions under different conditions, humans consciously capture the process of AI value judgement and provide feedback, and in the interactive relationship humans and AI can achieve value understanding and shift from unidirectionality to bidirectionality (Liu Wei, 2021[24]).

The mode of achieving value alignment on the basis of the human-machine interaction relationship is called interactive value alignment. It aims, in the course of realising human-machine interaction, for humans and AI to form a co-governing subject on the basis of their respective agency and of the interactivity between them (Gu Xinyi, 2023[25]), and to achieve value alignment through bidirectional adaptation. Compared with existing paths to value alignment, interactive value alignment has the following characteristics. First, the transparency of the alignment process. Interactive value alignment aims to achieve alignment in the form of simulations of human-machine cooperation and of social scenarios; in the course

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

of collaboration humans and machines appear explicitly and the interaction is directly visible, involving not only visual and auditory feedback but, with the development of embodied intelligence, also the interaction of bodily movement and facial expression (Song Chunyan, 2023[26]), and the alignment process is thereby made transparent. Second, the real-time character of value understanding. At the input end, human value feedback can be captured by AI in real time, and the objective data collected by sensors is integrated with human subjective perceptual information to form a new value identification. Unlike the mere embedding of moral norms, the information processed by AI combines the value norms already embedded with real-time value input to form a new value understanding. Third, the matching of value judgements. At the output end, AI matches the algorithmic output of computation and iteration against the value effects manifested in human decision-making to form a new judgement (Liu Wei, 2019[27]). Value alignment is realised within the interactive relationship. Just as the values of human society are formed over the long course of human intercourse, so in the creation of an interactive context the value judgements of machines will reach a higher degree of matching with human value judgements. Fourth, the plurality of interactive subjects. The subjects of interactive alignment encompass such plural agents as human-machine, society-machine and machine-machine; the processes of feedback and interaction allow many parties to participate jointly, can achieve the objective of aligning plural values, and facilitate the effective evaluation of the results of alignment. Interactive value alignment aims to construct an ecology of interaction among the human, machine and environmental systems, and to achieve the objective of value alignment effectively on the basis of objective data, subjective values and symbiotic interactive relations.

4.2 The implementation path of interactive value alignment

Interactive value alignment rests on a dynamic interactive process of bidirectional understanding between human users and AI, real-time feedback, the integration of information, value identification and effective judgement, so as to achieve value alignment between AI and humans. The implementation of interactive value alignment proceeds from the interactive subject status of AI and, on the basis of contextualised value consensus, achieves dynamic value alignment through simulations of human-machine cooperation and of social scenarios.

4.2.1 Granting AI interactive agency is the precondition of value alignment

In the course of human-machine interaction, AI possesses interactive agency. Interactive value alignment aims at a shift from the substantialist thinking of anthropocentrism to the relational cognitive thinking that no longer dwells on the dispute over whether AI possesses the status of a moral subject, but establishes its status as an interactive subject within the human-machine interaction relationship. The emergence of socially situated AI has further confirmed the interactive agency of AI (Krishna et al., 2022[28]): within a real social environment, an agent can achieve its own optimisation through interaction with humans and with the environment, and these forms of interaction are diverse — visual, linguistic, behavioural and so on. Using special sensors to capture tactile interaction between humans and AI, and machine-learning algorithms with C3D classification to identify tactile types, yields an average recognition rate as high as 95 per cent, which means that the real-time and sensory dimensions of human-machine interaction are technically achievable (Yang Qingfeng, 2023[29]). Diversified forms of interaction can further create possibilities for human-machine value alignment.

Interactive value alignment does not proceed from whether AI possesses ethical subject status or what degree of it, but regards AI as the other in the interactive relationship and as a moral enabler in the interactive process. Supported by massive data and complex algorithms, the AI that humans now face is no longer a single task-executor but an enabler at the foundational level. Owing to the complexity of algorithms in machine learning, machine cognition is not a simple superimposition and extension: the internal decisions of an algorithm change with deep learning on training data, and in this process a great many unpredictable processes and results even emerge. The cognitive and explanatory models of humans thus stand on one side of a gap in cognitive dimension from the complexity of algorithms in machine learning; the position of humans in cognitive activity has changed profoundly, and they are gradually losing the central position they occupied in epistemology. A non-anthropocentric epistemology must therefore be constructed, one that recognises the value machines ought to have in epistemology (Dong Chunyu, 2023[16]). Likewise, as a foundational enabler, AI supplies new value understanding and judgement through the process of interaction with humans. As Luisa Damiano's notion of synthetic ethics has it, against the background of continuing technological development humans can establish long-term, ecological and sustainable developmental relations with artificial agents (Damiano and Dumouchel, 2020[30]). In such relations, value alignment is the outcome of the symbiotic interaction between humans and AI; AI possesses the status of an interactive subject and can supply new understanding and content for the formulation of standards of appropriate behaviour. Interactive value alignment aims to make effective use of the background of AI technological development and to emphasise the alignment relation among actors in the context of human-machine interaction. Interactive value alignment requires AI, like human actors, to respond appropriately in appropriate social environments and situations. The standard of appropriateness is not pre-given and is not forcibly aligned with the standards of appropriate behaviour in human society; rather, the standard

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

of the rightness of behaviour is continually adjusted according to the specific situation and context.

4.2.2 Contextualised value consensus is the key to interactive value alignment

Just as intercourse among humans must conform to moral norms, the interaction between AI and humans likewise requires a common moral foundation, namely value consensus (Fairweather, 2004[31]). Every act of human intercourse may involve disagreement, yet consensus and reconciliation can still be reached, because human society possesses a common foundation of values. The key problem value alignment now faces is which values to align with and which principles to follow. Most existing paths to value alignment run into difficulty because of the uncertainty of values and their resistance to unification; interactive value alignment therefore attempts, starting from the interactive process, to take contextualised value consensus as the key to value alignment.

Contextualised value consensus focuses on the preconditions of value consensus and on the requirements of contextualisation. On the one hand, contextualised value consensus draws on the basic moral consensus of human society and supplies a moral framework. Just as procedural ethics establishes a moral framework stripped of content, value alignment still requires feedback on the appropriateness of concrete behaviour in the light of situational and relational realities. Concrete expressions of value consensus can be found in existing research: the RICE principles (robustness, interpretability, controllability and ethics), for example, specify the value alignment objectives of AI following human instructions, understanding human intentions and giving feedback on behavioural preferences in the course of action; the FATE principles (fairness, accountability, transparency and ethics) tend to define the higher-order values AI should possess in the human-machine interaction relationship; and the 3H standard (helpful, honest and harmless) aims to identify the most suitable principles of value alignment (Ji et al., 2024[15]). It is not difficult to see that these value principles embody the consistency of human values. The value consensus interactive value alignment aims to construct is the value consensus commonly observed in real human society: values that are macro in scope yet recognised and observed by virtually all human civilisations, such as the basic value principles of respect, justice, safety and beneficence. But just as consistent values in reality do not take the place of humans in ethical thinking and decision-making, so humans themselves form ethical consensus on the basis of common values in combination with real situations.

On the other hand, when contextualised value consensus is applied to concrete scenarios, the ordering of values is always contextualised, so value annotation must be carried out in the light of the specific situation in order to achieve bidirectional human-machine understanding and alignment. At the technical level, scholars have proposed the Axies model, which allows human users to participate in the process by which machines understand value consensus. The model uses value-laden corpora supplied by human users and natural language processing technology to guide AI systems in identifying the values of a particular context (that is, the linguistic situation) and in carrying out value annotation, on which human users then give feedback. This model transforms the abstract task of value identification into the concrete task of human-machine interactive value annotation (Liscio et al., 2022[32]) and, by incorporating the linguistic situation, confirms AI's understanding of value consensus in bidirectional human-machine interaction and feedback. Contextualised value consensus is the key to ensuring that AI systems observe moral standards consistent with those of humans when performing tasks or assisting human decision-making, but its definition and evaluation are themselves challenging. Interactive value alignment therefore also requires the further enrichment of the content of moral consensus and the construction of corresponding value datasets for training. Denis Emelin and colleagues, for example, introduced the "Moral Stories", "Social Chemistry" and "moral corpus" datasets, which supply basic learning data on the moral norms of human society (Emelin et al., 2020[33]); datasets such as "Commonsense Morality question answering" are used to test the capacity of models to align morally with human values (Hendrycks et al., 2020[34]); and the "Commonsense Norm Bank" is used to further complete AI's training in human moral annotation (Jiang et al., 2022[35]). Because value datasets are static in nature, they contain only relatively fixed content for value assessment and are also vulnerable to interference from targeted training; contextualised value consensus therefore still requires the simulation of concrete scenarios in order to achieve dynamic alignment.

4.2.3 Achieving dynamic alignment through human-machine cooperation and social scenario simulation

Interactive value alignment emphasises the processual and relational character of alignment, and aims to achieve dynamic alignment through simulations of human-machine cooperation and of social scenarios. On the one hand, human users interact with AI by supplying value-laden viewpoints grounded in social values; on the other hand, AI, on the basis of contextualised value consensus, completes the annotation and understanding of values within the scenario and is optimised through human feedback and evaluation. In this process moral design neither simply inputs fixed value principles into intelligent machines nor has humans stand outside and supply value feedback: alignment is achieved in dynamic human-machine interaction. AI becomes a participant in and enabler of the value process within a dynamic relationship, continually adjusting and improving its understanding of values in concrete interactive situations.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Human-machine cooperation simulation for alignment. Simulating human-machine cooperation restores the process of human cooperation: humans can grasp, through the "behaviour" of an intelligent robot, a definite meaning similar to that of human subject behaviour, and can form stable expectations of the consistency of its "behavioural" responses (Liu Zhe, 2022[36]). In this process the intelligent robot is no longer an object saturated with complexity and unfamiliarity; the controllability of AI is thereby achieved and the objective of value alignment realised. A number of technical methods already attempt, by simulating human-machine cooperation, to enable AI to learn human values. The "XAI" system designed by Yuan Luyao and her team, for example, provides a method of bidirectional human-machine value alignment. First, the system starts from bidirectional communication between humans and AI: the AI must extract useful information from human feedback in order to infer user values and adjust its strategies accordingly. Second, the AI is required to explain its decision process to the user and to predict the user's values on the basis of on-the-spot feedback. Third, the user is empowered to check whether human values have been shared in the course of cooperation, realising a human-centred approach, driving dynamic machine-human and dynamic human-human exchange, and guiding the machine to absorb the user's values (Yuan et al., 2022[37]). In this process the traditional data-driven mode of machine learning is replaced by communicative learning within cooperation; the task process simulates cooperation in the real human social world and takes value learning as its task objective, extending the effective methods of human-machine interactive value alignment. Through the form of human-machine cooperation, AI can learn and infer the values of human users, human users can evaluate and understand the value output of AI, and the objective of value alignment is effectively achieved in the bidirectional process of cultivating human-machine value understanding.

Social scenario simulation for alignment. Interactive value alignment creates simulations of social scenarios in which, through interaction with the environment, AI gradually acquires situational sensitivity. Luciano Floridi has argued that a moral outlook is founded on sufficient interactivity, autonomy and adaptability, and on ways of behaving that conform to moral standards (Floridi and Sanders, 2004[38]). In the intelligent era of human-machine integration, AI needs to be trained in interaction to understand the responsibilities of other moral subjects and to cultivate the capacity to act for the sustainable and sound development of human society. Through social scenario simulation, AI can achieve formalised cognition of human behaviour and social structure, contextualised understanding, and the reasonable output of "behaviour", thereby achieving the objective of dynamic value alignment. Scholars have already constructed models of real society, created plural contexts and simulated concrete social scenarios in order to achieve alignment. Liu Ruibo and colleagues, for example, established the "Stable Alignment" algorithm, which places AI in a simulated human social sandbox and, by imitating human social interaction, allows AI to learn the plural value orientations of human society (Liu et al., 2023[39]). The human social sandbox is composed of social agents represented by multi-language models; these agents interact with one another and record their behaviour, presenting typical instances of alignment and misalignment in the interaction data. The researchers further subject these interaction data to collective rating and detailed feedback in order to revise the response mechanism, thereby progressively improving responses in each round of simulation and enhancing the degree of alignment, which supplies an effective path for interactive value alignment.

Humans learn social norms, determine value orientations and adjust their behaviour through cooperation and situated interaction. For AI, a language model is by nature trained in social isolation: it cannot directly experience real social life, nor can it obtain multi-round feedback from others in order to improve. Interactive value alignment aims to overcome the existing modes of technical training and ethical design by creating a relatively realistic context through simulations of human-machine cooperation and of social scenarios, and by further improving the quality of value alignment in AI models through multi-round interactive evaluation.

4.3 The evaluation of interactive value alignment

Interactive value alignment does not entirely negate or overturn the existing paths to value alignment. On the basis of establishing the interactive subject status of AI and contextualised value consensus, it establishes a long-term and sustainable developmental relationship between humans and AI through simulations of human-machine cooperation and of social scenarios. Proceeding from relational theory, interactive value alignment restores the appropriateness of AI behaviour to the concrete process of interaction between humans and AI, and value alignment becomes the product of the human-machine interactive relationship. The value alignment that the development of AI requires does not emphasise only human obligations or machine responsibilities; it treats alignment as a value category capable of adjusting the relation between the two. Interactive value alignment therefore regards the human-machine relationship as a partnership of complementary strengths rather than as an alignment subject split between subject and object, so as to achieve the objective of value alignment founded on "human-machine symbiosis".

What merits further exploration is that the future of interactive value alignment requires the coordinated development of such technologies as socially situated AI, situational robotics and embodied intelligence. The reason is that the key to interactive value alignment lies in the construction of the interaction relationship between AI and humans, and the basis for

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

its consideration and evaluation lies in how AI performs within that relationship. The differences in the performance of AI and humans within the interactive relationship arise from differences in sociality, corporeality and affectivity. As Stuart Armstrong puts it, AI does not take human goals as its own; it is not an organism, and it may not even be able to understand the pain humans experience (Armstrong, 2014[40]). The human brain differs from AI algorithms: thought is not carried on through symbol processing, and values also involve implicit knowledge — knowledge founded on perception rather than on knowing. AI cannot capture this background meaning and knowledge; only humans can understand the relational meaning within it, because humans are concrete, real beings situated in the world. Technologies of embodied, situated and social AI can therefore more effectively achieve the objective of interactive value alignment.

5 Conclusion

The objective of AI value alignment is to make large models more reliable, practical and safe tools, and it bears on the future development of general-purpose AI. Most existing paths to value alignment start from technical means, but aligning only on the basis of human instructions and preference feedback ignores the importance of value content and dynamic context, and lacks the intrinsic knowledge and generative background concerning what genuinely "good" behaviour is. To promote consistency between the development of AI and human value objectives and to make AI better conform to human expectations, deeper understandings of value theory, interactive relations and social context must be incorporated into the process of value alignment. By establishing AI as an interactive subject and as an "enabler" of aligned values, and on the basis of contextualised value consensus, interactive value alignment achieves processual and relational dynamic alignment through simulations of human-machine cooperation and of social scenarios, ensuring that the behaviour and decisions of large models are continually optimised in human-machine interaction and can conform to the value expectations of human society.

Funding

This work was supported by the Special Fund for Basic Scientific Research Operating Expenses of Central Universities (2662023MYQD001) and the Huazhong Agricultural University Postgraduate Training Conditions Construction Project (2023KC06).

References

[1] Scheutz, M. The inherent dangers of unidirectional emotional bonds between humans and social robots[M]//Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge: The MIT Press, 2012.

[2] Government of China. Interim Measures for the Administration of Generative Artificial Intelligence Services[R/OL]. (2023-07-10)[2024-07-10]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[3] Lanier, J. The myth of AI[EB/OL]. (2014-11-14)[2024-07-10]. https://www.edge.org/conversation/the-myth-of-ai#26015.

[4] Wiener, N. Some moral and technical consequences of automation: as machines learn they may develop unforeseen strategies at rates that baffle their programmers[J]. Science, 1960, 131(3410).

[5] Leike, J., Krueger, D., Everitt, T., et al. Scalable agent alignment via reward modeling: a research direction[EB/OL]. (2018-11-09)[2024-07-10]. https://arxiv.org/abs/1811.07871.

[6] Gabriel, I. Artificial intelligence, values, and alignment[J]. Minds and Machines, 2020(30).

[7] Tencent Research Institute. Value alignment of large AI models: what, why and how?[EB/OL]. (2023-08-24)[2024-07-10]. https://www.tisi.org/26547.

[8] Wu Guanjun. The trust problem of large language models and the logic of capital[J]. Contemporary World and Socialism, 2023(5).

[9] Anthropic. Claude's constitution[EB/OL]. (2023-05-09)[2024-07-10]. https://www.anthropic.com/index/claudes-constitution.

[10] Shevlane, T., Farquhar, S., Garfinkel, B., et al. Model evaluation for extreme risks[EB/OL]. (2023-09-22)[2024-07-10]. https://arxiv.org/abs/2305.15324.

[11] Bakker, M. A., Chadwick, M. J., Sheahan, H. R., et al. Fine-tuning language models to find agreement among humans with diverse preferences[EB/OL]. (2022-11-28)[2024-07-10]. https://arxiv.org/abs/2211.15006.

[12] Yu Xue. Approaches to the moral design of intelligent machines and their attribution of responsibility[J]. Studies in Ethics, 2022(4).

[13] Gong Qun. On the moral examination of weak artificial agents[J]. Philosophical Research, 2023(3).

[14] Coeckelbergh, M. Moral appearances: emotions, robots, and human morality[J]. Ethics and Information Technology, 2010(12).

[15] Ji, J. M., Qiu, T. Y., Chen, B. Y., et al. AI alignment: a comprehensive survey[EB/OL]. (2024-02-27)[2024-07-10]. https://alignmentsurvey.com.

[16] Dong Chunyu. The nature and limits of artificial intelligence seen from the opacity of machine cognition[J]. Social Sciences in China, 2023(5).

[17] Aristotle. Nicomachean Ethics[M]. Trans. Liao Shenbai. Beijing: The Commercial Press, 2016.

[18] Kenward, B., Sinclair, T. Machine morality, moral progress, and the looming environmental disaster[J]. Cognitive Computation and Systems, 2021.

[19] Hu Shenglan. An analysis of the moral empowerment of artificial affective agents[J]. Studies in Dialectics of Nature, 2023, 39(2).

[20] Rosalind Picard. Affective Computing[M]. Trans. Luo Senlin. Beijing: Beijing Institute of Technology Press, 2005.

[21] Wu Tongli. Is artificial intelligence qualified to become a moral subject?[J]. Philosophical Trends, 2021(6).

[22] Fu Changzhen. Will robots have "empathy"? A perspective based on Confucian affective ethics[J]. Philosophical Analysis, 2019, 10(6).

[23] Sun Weiping. Algorithmic discrimination and social justice from the perspective of value philosophy[J]. Philosophical Research, 2023(3).

[24] Liu Wei. Human-Machine Integration: Beyond Artificial Intelligence[M]. Beijing: Tsinghua University Press, 2021.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

[25] Gu Xinyi. Interactive autonomy and its ethical implications under brain-computer integration[J]. Journal of Dialectics of Nature, 2023, 45(7).

[26] Song Chunyan. Self-consciousness and interactive subjectivity in human-machine integrated intelligence[J]. Studies in Ethics, 2023(5).

[27] Liu Wei. The present state and prospects of human-machine integrated intelligence[J]. National Governance, 2019(4).

[28] Krishna, R., Lee, D., Li, F.-F., et al. Socially situated artificial intelligence enables learning from human interaction[EB/OL]. (2022-06-14)[2024-07-10]. https://www.pnas.org/doi/epdf/10.1073/pnas.2115730119.

[29] Yang Qingfeng. The myth of artificial intelligence, superintelligence and its contract ethics[J]. Journal of Shanxi University (Philosophy and Social Sciences Edition), 2023, 46(6).

[30] Damiano, L., Dumouchel, P. Emotions in relation: epistemological and ethical scaffolding for mixed human-robot social ecologies[J]. HUMANA. MENTE Journal of Philosophical Studies, 2020, 13(37).

[31] Fairweather, N. B. Why incomplete codes of ethics are worse than none at all[M]//Computer Ethics and Professional Responsibility. Malden: Blackwell Publishing, 2004.

[32] Liscio, E., Meer, M. V. D., Siebert, L. C., et al. What values should an agent align with? An empirical comparison of general and context-specific values[J]. Autonomous Agents and Multi-Agent Systems, 2022, 36(23).

[33] Emelin, D., Bras, R. L., Hwang, J. D., et al. Moral stories: situated reasoning about norms, intents, actions, and their consequences[EB/OL]. (2020-12-31)[2024-07-10]. https://arxiv.org/pdf/2012.15738v1.

[34] Hendrycks, D., Burns, C., Basart, S., et al. Aligning AI with shared human values[EB/OL]. (2020-08-05)[2024-07-10]. https://arxiv.org/pdf/2008.02275v1.

[35] Jiang, L. W., Hwang, J., Bhagavatula, C., et al. Can machines learn morality? The Delphi experiment[EB/OL]. (2022-07-12)[2024-07-10]. https://arxiv.org/abs/2110.07574.

[36] Liu Zhe. The hidden worry of bodily alienation in the age of artificial intelligence: reflecting on the interactive relationship between humans and intelligent robots from a phenomenological perspective[J]. Foreign Philosophy, 2022(2).

[37] Yuan, L. Y., Gao, X. F., Zheng, Z. L., et al. In situ bidirectional human-robot value alignment[J]. Science Robotics, 2022(7).

[38] Floridi, L., Sanders, J. W. On the morality of artificial agents[J]. Minds and Machines, 2004, 14(3).

[39] Liu, R., Yang, R. X., Jia, C. Y., et al. Training socially aligned language models on simulated social interactions[EB/OL]. (2023-10-28)[2024-07-10]. https://arxiv.org/pdf/2305.16960.

[40] Armstrong, S. Smarter than Us: The Rise of Machine Intelligence[M]. Berkeley: Machine Intelligence Research Institute, 2014.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.