

# The Appropriateness and Boundary Governance of Embedding Artificial Intelligence Technology into College Students' Value Guidance

Jing Ren<sup>1</sup>, Xiaoguang Sun<sup>2\*</sup>

<sup>1</sup>Inner Mongolia Technical College of Mechanics and Electrics, Hohhot, 010070, China

<sup>2</sup>Inner Mongolia University of Finance and Economics, Hohhot, 010070, China.

\*Corresponding author: sunxg@imufe.edu.cn

**Abstract:** As artificial intelligence technology permeates the fields of information distribution and cognitive interaction, it has been introduced into the domain of guiding college students' core socialist values, becoming a new variable affecting the value shaping of young people. However, technological embedding is not value-neutral; while improving the efficiency of information delivery, it may trigger risks such as digital alienation of the guidance subject, impairment of value autonomy, and the usurpation of value rationality by instrumental rationality. Grounded in the perspective of Marxist theory, this study elucidates the logic of appropriateness from theoretical thresholds such as the compatibility between technological functions and human subjectivity, the structural coupling of mediating effects and value goals, and the free and comprehensive development of human beings. It analyzes boundary tensions including the substitution of guiding subjectivity, algorithmic cognitive discipline, and conflicts with value autonomy. Furthermore, it proposes governance mechanisms such as context-based hierarchical regulation, explainable constraints framed by core socialist values, and division of responsibilities and feedback calibration in human-machine collaboration. These provide a theoretical boundary framework for the moderate embedding of artificial intelligence in the guidance of core socialist values, ultimately aiming at the fundamental goal of cultivating new generations capable of undertaking the mission of national rejuvenation.

**Keywords:** artificial intelligence technology embedding; value guidance; appropriateness; boundary tension; governance mechanism

## Introduction

The guidance of college students' values takes the cognitive internalization of value norms and the consistency of behavioral performance as its core objectives, which heavily depends on the holistic grasp of individuals' cognitive states, emotional tendencies, and contextual characteristics. With its capabilities in information filtering, preference prediction, and semantic understanding, artificial intelligence technology offers new technological possibilities for differentiated provision and dynamic feedback within the guidance field. However, there exists an inherent tension between the computable logic of technological systems and the nonlinear, context-dependent nature of value guidance. Unbounded technological intervention may give rise to risks such as the replacement of guidance professionals' expert judgment by algorithmic recommendations, the implicit disciplining of individual value autonomy, and the distortion of guidance objectives during quantitative translation. Given this, it is theoretically necessary to explore the logic of appropriateness and boundary governance regarding the embedding of artificial intelligence into value guidance, which requires specifying the conditions and limits of technological adaptation while simultaneously constructing a governance framework that responds to the risks of embedding. This study proceeds from three dimensions — appropriateness judgment, boundary tension identification, and governance mechanism design — to provide an analytical pathway for the coordinated coexistence between technological embedding and the intrinsic logic of guidance.

## **1. The Logic of Appropriateness for Embedding Artificial Intelligence Technology into College Students' Value Guidance**

### ***1.1 Functional Adaptation of Artificial Intelligence Technology in the Guidance Field***

The functional adaptation of artificial intelligence technology in the field of guiding core socialist values is first manifested as the restructuring of information flow architecture and cognitive exposure frequency. Recommendation algorithms based on collaborative filtering and content tag matching generate differentiated information sequences according to individual interaction patterns, thereby transforming the transmission of value symbols from one-way broadcasting to multi-node distribution; the programmability of technology allows multimodal content to be presented with preset weights, achieving dynamic coordination between guidance objectives and cognitive inertia in information delivery. This adaptation goes beyond efficiency improvement; it reshapes the diffusion topology of value information at the group level, enabling the differentiated allocation of guidance resources in accordance with individuals' cognitive states, and demonstrates the positive potential of technology as an extension of human's essential powers<sup>[1]</sup>.

From the perspective of the depth of functional adaptation, natural language processing and affective computing provide an operationalizable pathway for semantic understanding and emotion recognition. By quantifying the attitudinal tendencies and emotional polarities in individual speech, the system can identify potential deviations in value cognition under non-interventional conditions, such as detecting signs of alienation from patriotism, dedication, integrity, and friendliness. Such identification does not replace the professional judgment of guidance practitioners, but rather serves as an auxiliary cognitive tool to reduce blind spots in guidance caused by information asymmetry. Technological adaptability is thus manifested as multiple mappings of information flow, emotional flow, and value flow, and its fundamental criterion lies in whether it maintains and enhances the subject position of human beings, so that technology genuinely serves the refined needs of promoting the comprehensive development of individuals.

### ***1.2 The Coupling of Value Guidance Objectives and Technological Mediation Effects***

The guidance of core socialist values aims to promote the cognitive internalization and behavioral externalization of individual-level value goals, social-level value orientations, and personal-level value norms among college students. When artificial intelligence intervenes as a mediating variable, its technological mediation effect is manifested as the recoding of information filtering, ordering, and presentation modes, with algorithmic parameters directly determining the visibility gradient of value-related information. Such mediation is not a value-neutral technical operation, but rather transforms guidance objectives into a computable logic of weight allocation. The prerequisite for preventing instrumental rationality from usurping value rationality lies in whether the algorithmic parameters can bear the directional requirements entailed by value guidance.

The establishment of coupling relationships must satisfy the condition that there exists a structurally mappable similarity between the parameter settings of technological mediation and the value orientations of guidance objectives. Specifically, the value hierarchy advocated by core socialist values needs to be embedded into the loss functions or reward mechanisms of algorithms through quantifiable indicators; however, such quantification must not come at the cost of dissolving the rich connotations of values. For example, diversity metrics of value information can be designed as regularization terms in recommendation systems to curb the excessive accumulation of homogeneous content and ensure that pluralistic values such as justice and harmony are presented in a balanced manner. When the technological mediation effect can stably output semantic biases consistent with specific value goals such as patriotism, dedication, and integrity, a functional coupling is formed between the two. This coupling does not eliminate the multiple possibilities of individual value choices; rather, it enables the technological system to maintain structural channels for the continuous infiltration of core socialist values within its parameter space, while simultaneously allowing individuals to conduct autonomous exploration within the reasonable information boundaries delineated by the system, thereby accumulating subjective strength for the advancement toward free and comprehensive development.

### ***1.3 Multidimensional Theoretical Foundations and Threshold Delineation for Appropriateness Judgment***

The judgment of appropriateness needs to draw its foundations from the Marxist critique of

technological alienation, the philosophy of human subjectivity, and the theory of the free and comprehensive development of human beings. From the perspective of philosophy of technology, artificial intelligence, as a tool for cognitive extension, is appropriate only if it enhances rather than replaces the interpretive authority of guidance practitioners, thereby avoiding the alienated state in which technology enslaves human beings; from the perspective of philosophy of subjectivity, technological embedding must not undermine metacognitive monitoring or weaken the capacity for autonomous choice; and from the perspective of the theory of comprehensive development, technology must not narrow the process of value guidance or weaken emotional bonds. These three dimensions constitute a theoretical triangle, with the ultimate criterion being whether they are conducive to cultivating new generations of the era who possess independent personality and sound value judgment.

Based on the above theoretical foundations, we can establish three-tiered threshold intervals for appropriateness. The primary threshold concerns whether technological intervention causes semantic distortion of guidance objectives, that is, whether the interpretability of value information after algorithmic processing falls below an acceptable lower limit, resulting in the misinterpretation of the core connotations of core socialist values. The secondary threshold evaluates the degree to which the technological system compresses the space for professional judgment of guidance practitioners; if the coerciveness of algorithmic recommendations exceeds the veto authority of the guidance practitioners, it indicates that the technology has shown signs of transgression and alienation. The tertiary threshold measures the level of impairment to individual value autonomy; when individuals are unable to provide meaningful retrospective explanations for the reasons behind their own value choices, it signifies that technological embedding has exceeded the appropriate boundary and begun to deviate from the direction of the free and comprehensive development of human beings. Threshold delineation does not consist of fixed numerical values, but rather serves as a boundary reference system that dynamically adjusts with technological iteration and changes in guidance contexts; its calibration frequency should match the speed of system behavior drift, so as to continuously safeguard the subject position of human beings<sup>[2-5]</sup>.

## **2. Boundary Tensions of Embedding Artificial Intelligence Technology into College Students' Value Guidance**

### ***2.1 Potential Mechanisms of Technological Intrusion in Undermining Guiding Subjectivity***

The erosion of guiding subjectivity by technological intrusion first manifests as the substitution of subject functions in the process of decision-making delegation. When artificial intelligence systems assume functions such as information filtering, preference prediction, and path planning in the context of value guidance, the active judgment and contextual evaluation of guidance practitioners are gradually overridden by the standardized recommendations output by algorithms. This substitution is not an explicit transfer of authority, but rather an implicit path dependence formed in daily guidance operations through the reduction of decision-making costs. Guidance practitioners tend to adopt the content strategies and interaction methods recommended by the system, thereby weakening their own differentiated judgments based on professional experience and relational sensitivity; guiding subjectivity is thus displaced by the presupposed logic of the technological system.

Another dimension of the undermining mechanism stems from the exclusion of guidance practitioners' reflective control by the black-box characteristics of technological systems. The parameter scale and multi-layer nonlinear transformations of artificial intelligence models make it difficult to fully trace the causal chain between inputs and outputs. Guidance practitioners are unable to provide adequate explanations for why the system recommends a particular sequence of value information, and thus their supervisory and corrective functions are constrained. Under the premise that they cannot confirm whether the technological output aligns with the guidance intentions, the regulatory capacity of guidance practitioners is suspended, and subjectivity gradually yields to the autonomous operation of the technological system. This erosion is not accomplished instantaneously, but rather constitutes a structural deviation that accumulates cumulatively in the process of repeatedly invoking technological tools.

### ***2.2 Analysis of the Conflict between Algorithmic Cognitive Discipline and Value Autonomy***

The basic operational mode of algorithmic cognitive discipline is to shape individuals' attention allocation and memory consolidation regarding specific value information through repeated exposure

and reinforcement through feedback. The information flow generated by recommendation systems based on individuals' historical behaviors continuously converges toward the same semantic direction of values, forming a flexible cognitive guidance mechanism. Under the effect of this mechanism, the range of value options accessible to individuals is systematically narrowed, and the multi-perspective reference framework required for value judgment becomes distorted. The core of value autonomy lies in the individual's ability to compare, weigh, and make conscious choices among multiple value options; however, algorithmic discipline precisely reduces the space for comparison by diminishing option diversity, thereby generating a structural conflict with value autonomy<sup>[6]</sup>.

The deeper manifestation of this conflict is also reflected in individuals' confusion about the origins of their own value preferences. Those who have been immersed in an environment of algorithmic discipline for an extended period often struggle to distinguish whether their value inclinations stem from genuine autonomous reflection or from passive accumulation shaped by repeated system recommendations. When individuals misperceive the cognitive inertia induced by the technological system as an extension of their own volition, value autonomy undergoes an unnoticed yet fundamental shift. This conflict cannot be resolved by simply deactivating algorithmic recommendations, because it has already become entrenched in individuals' information-acquisition habits and evaluative frameworks. At its core, the conflict arises from an incommensurable logical divergence between the system's objective function—typically maximizing user engagement or informational consistency—and the pluralistic openness that value autonomy necessarily demands.

### ***2.3 Manifestations of Boundary Disorder Induced by the Computability of the Guidance Process***

The computability of the guidance process first manifests as inaccuracies in the quantitative translation of value objectives. Dimensions such as emotional identification, meaning construction, and value internalization involved in value guidance are non-discrete and nonlinear in nature, and thus cannot be fully mapped into computable indicator variables. When artificial intelligence systems decompose guidance objectives into proxy indicators such as click-through rates, dwell time, and sentiment polarity scores, a semantic gap inevitably exists between the measured targets and the substantive objectives. This translation inaccuracy causes the technological system to optimize the measurable proxy variables rather than the effectiveness of guidance itself, thereby triggering a boundary disorder in which instrumental rationality deviates from value rationality within the guidance process.

Another manifestation of boundary disorder is the conflict between the automated decision-making of the technological system and the specificity of the guidance context. Value guidance heavily depends on the holistic grasp of individuals' immediate states, relational atmospheres, and contextual information, whereas artificial intelligence systems can only process structured features that have already been encoded. When the technological system outputs guidance recommendations based on limited features, its decision-making logic may severely deviate from the appropriate path for that particular context. If guidance practitioners follow the system's recommendations, the outcomes may contradict the original intent of guidance; if they override the system's suggestions, the value of technological intervention is suspended. This dilemma signifies that the boundary of the guidance process has shifted from functional assistance to functional interference, forming an irreconcilable zone of tension between the operational logic of the system and the intrinsic logic of guidance<sup>[7]</sup>.

## **3. Boundary Governance Mechanisms for Risks of Inappropriate Embedding**

### ***3.1 Hierarchical Regulation of Technological Intervention Based on Guidance Contexts***

The hierarchical regulation of technological intervention sets functional boundaries of the system according to differences in guidance contexts, with the fundamental principle of safeguarding ideological security and the bottom line of human subjectivity. Contexts are categorized by information sensitivity, autonomy requirements, and depth of emotional involvement into levels such as low-sensitivity information transmission, attitude negotiation, and value conflict reconciliation: the low-sensitivity level may undertake information retrieval and aggregation; the attitude negotiation level only provides multi-path options while reserving the final judgment authority of guidance practitioners; the value conflict reconciliation level is limited to emotional recognition assistance and prohibits the generation of prescriptive suggestions with value biases. The hierarchical division is based on the degree of dependence of guidance objectives on the subjective judgment of educators; the higher the

dependence, the lower the authority of technological intervention, so as to prevent the overstepping of instrumental rationality.

The implementation of hierarchical regulation requires establishing an operational indicator system for context classification, integrating the experience of ideological work in higher education institutions. The indicator system comprises three dimensions: value cognitive stability, degree of value consensus, and density of emotional interaction. It is particularly crucial that, for ideological issues involving major matters of principle, algorithms shall not autonomously generate value judgment content; they can only push authoritative information that has been strictly manually reviewed, with the system degenerating into an information transmission channel. When a task falls into the high-risk level, the system functions should degenerate into passive monitoring. Hierarchical regulation does not comprehensively suppress technological functions; rather, through context-sensitive tiered authorization, it aligns the intensity of intervention with the level of ideological risk, ensuring that technology serves the fundamental task of fostering virtue and cultivating talent.

### ***3.2 Value-Oriented Constraint Principle of Technological Explainability***

The constraint principle of technological explainability requires that when the system is embedded in the guidance of core socialist values, the decision-making basis and output logic must be comprehensible and traceable to guidance practitioners, and must be presented within the discourse framework of core socialist values. This principle differs from general algorithmic explainability, which only focuses on parameter visualization, and further requires that the causal relationship between output information and value objectives be presented in a semanticized manner. That is, when the system recommends content, it must simultaneously output the judgment path of value relevance and explain it in a discourse consistent with the value requirements of patriotism, dedication, and integrity, rather than merely providing a list of feature weights. This enables guidance practitioners to clearly judge the intrinsic consistency between the system's value orientation and core socialist values.

The specific implementation of the constraint principle can be achieved through two technological pathways: attention mechanism visualization and counterfactual explanation. Attention mechanism visualization demonstrates which semantic units are assigned higher weights by the model when processing input information, thereby revealing the basis for screening value-related information; counterfactual explanation, through comparative analysis, presents the necessary conditional relationships between value elements and system decisions. Both types of explanations ultimately need to be transformed into natural language summaries with core socialist values as the basic vocabulary, for example, stating "this content is recommended because it embodies the values of justice and friendliness, while certain content is suppressed because it presents a risk of contradicting integrity," rather than outputting technical charts or mathematical symbols. The bottom-line requirement for explainability constraints is that guidance practitioners must be able to complete a reasonable assessment of the system's value-judgment logic under limited cognitive load and intervene promptly upon detecting deviations. Any technological system that fails to meet this bottom-line requirement, regardless of the superiority of its algorithmic performance, should not be deployed in value guidance scenarios, so as to prevent technological logic from overriding human value judgment<sup>[8-10]</sup>.

### ***3.3 Division of Responsibilities and Feedback Calibration in Human-Machine Collaborative Guidance***

The framework for division of responsibilities needs to reflect the organizational characteristics of higher education institutions, based on distinguishing responsibilities in algorithm design, system deployment, and guidance implementation. Responsibility for algorithm design rests with the development party, which includes embedding value-sensitive design and bias detection; responsibility for system deployment rests with the introducing organization, covering context-adaptability assessment and hierarchical regulation setting, with ultimate responsibility for ideological gatekeeping borne by the leadership body for ideological work in the institution; responsibility for guidance implementation rests with frontline ideological and political educators, whose core duty is the prudent review and final decision-making regarding technological outputs. These three parties form a chain-like structure; the absence of any link will lead to a rupture in accountability, creating an institutional gap for digital alienation.

The feedback calibration mechanism aims to address the issue of behavioral drift caused by the dynamic evolution of technological systems, and embodies the characteristics of collaborative

governance involving multiple stakeholders. As data distribution changes or models are updated, the initial hierarchical regulations and explainability constraints may gradually become ineffective; therefore, it is necessary to establish a bidirectional regulatory loop: the usage experiences and judgment outcomes of guidance practitioners are periodically encoded as structured feedback and fed into the constraint parameter adjustment module; meanwhile, changes in system behavior are fed back in real time through monitoring indicators, triggering manual review when necessary. A joint mechanism should be established, in which ideological and political course teachers, counselors, and student leaders regularly conduct joint assessments of ideological trends, incorporating the practical wisdom of frontline educators and the authentic feedback of students into the calibration loop, so as to prevent the superposition and amplification of individual cognitive biases and technological system deviations through collective assessment. The calibration cycle is dynamically triggered based on the threshold of behavioral deviation, rather than at fixed intervals. Through the coordinated operation of a static framework of responsibility division and dynamic calibration involving multiple stakeholders, human-machine guidance can be maintained within appropriate boundaries, enabling artificial intelligence to serve as an auxiliary force for promoting the comprehensive development of human beings, rather than an alien and dominating force.

## Conclusion

From the perspective of Marxist theory, this study has constructed an analytical framework for the embedding of artificial intelligence into the guidance of college students' core socialist values, encompassing three dimensions: the logic of appropriateness, boundary tensions, and governance mechanisms. At the level of appropriateness, it has revealed that functional adaptation of technology must be conditioned upon the enhancement of human subjectivity and take the free and comprehensive development of human beings as its ultimate yardstick; at the level of boundary tensions, it has elucidated risks such as digital alienation, algorithmic discipline, and the usurpation of value rationality by instrumental rationality, the root cause of which lies in the incommensurability between the objective function of technological systems and the logic of comprehensive human development; and at the level of governance mechanisms, it has proposed three pathways: context-sensitive hierarchical regulation, explainability constraints framed by the discourse of core values, and division of responsibilities with feedback calibration. The deeper inquiry of this study concerns what kind of technological use can truly serve the cultivation of new generations capable of undertaking the mission of national rejuvenation. Future research needs to advance adaptive calibration of appropriateness thresholds, differentiated response studies for various value dimensions, construction of responsibility attribution models grounded in Marxist subjectivity theory, and comparative analyses of alienation risk differences across different technological architectures, so that artificial intelligence can genuinely become an auxiliary force for the comprehensive development of human beings.

## Fund Projects

The research project titled "Research on the Precision Teaching Model of Human-Machine Collaboration for the 'Corporate Culture' Course in the Context of Digital Transformation" under the Educational Science Planning Project of the Inner Mongolia Autonomous Region has been concluded, with project number: NGJGH2024195.

## References

- [1] Huang Xinrong, and Li Run. "Experimental Ideological and Political Education: Improvement and Innovation of the Disciplinary System of Ideological and Political Education in the Age of Digital Intelligence." *Ideological and Theoretical Education* 2026(3): 52-58.
- [2] Shi Yunxia, and Zhou Jinxin. "A Study on the Innovation of Ideological and Theoretical Education of the Communist Party of China in the New Era." *Guide to Ideological and Theoretical Education* 2026(3): 134-143.
- [3] Feng Gang, and Yu Yifei. "The Disciplinary Significance of Ideological and Political Work: Theoretical Elucidation, Empirical Reflections, and Development Prospects." *Ideological and Political Education Research* 42(1): 1-7.
- [4] Ai Jun. "Discourse Innovation and Path Optimization of Ideological and Political Education in Colleges and Universities under the Background of Youth Subculture." *Ideological Education Research*

2026(2): 138-143.

[5] Xu Rong. "Some Reflections on the Construction of the Autonomous Knowledge System of the Discipline of Ideological and Political Education in the Age of Digital Intelligence." *Research on Marxist Theory Disciplines* 11(10): 94-103.

[6] Dai Yuqi. "The Leading Force of Ideological and Political Education: A Signature Concept of the Discipline of Ideological and Political Education." *Research on Marxist Theory Disciplines* 11(8): 105-112.

[7] Zeng Linghui. "On the Innovative Development of the Basic Theories of the Discipline of Ideological and Political Education in the New Era." *Ideological and Theoretical Education* 2025(3): 49-56.

[8] Zhou Ke. "The Contemporary Interpretation of Marx's Theory of Alienation and Its Significance." *Theoretical Exploration* 2025(1): 13-21.

[9] Shen Guangming. "The Ecological Narrative of Marxist Theory of Productive Forces." *Marxist Studies* 2025(11): 67-74.

[10] Zeng Ruiming. "The Scientific Clarification of the Concept of Marxist Social Revolution and Its Theoretical Effects." *Marxist Studies* 2025(1): 41-51.