

Algorithmic Black Box and Value Bias: Ethical Risk Scrutiny and Safety Governance of Generative AI Embedding in Content Production of Internet Ideological and Political Education in Universities

Yuyang Xu ¹, Feng Zhong ², Yingmei Li ^{3,*}

¹ School of Data Science and Artificial Intelligence, Guangdong University of Finance, Guangzhou, China

² School of Credit Management, Guangdong University of Finance, Guangzhou, China

³ Qingyuan Campus, Guangdong University of Finance, Guangzhou, China

ABSTRACT

With the rapid development of Generative Artificial Intelligence (AIGC) technologies and their deep embedding into the entire process of content production for Internet ideological and political education in universities, their powerful automated generation and multi-modal interaction capabilities have empowered educational innovation while simultaneously introducing deep ethical risks and security challenges. The inherent "algorithmic black box" attribute of Generative AI leads to the uninterpretability of content production and the concealment of responsible subjects. Meanwhile, potential biases in training data can easily trigger "value bias" in output content, manifesting as the implicit penetration of ideology, the breeding of historical nihilism, and the obscuration of mainstream values. Based on the perspective of the overall national security concept and Marxist technological ethics, this paper deeply scrutinizes the multi-dimensional ethical dilemmas caused by the embedding of Generative AI into university online ideological and political content production, such as data privacy leakage, the dissolution of teacher-student subjectivity, and the solidification of algorithmic discrimination. On this basis, the article systematically constructs a full-chain safety governance system covering "data source purification—algorithmic model alignment—human-machine collaborative review—dynamic risk monitoring" from four dimensions: technical governance, subject reshaping, institutional standardization, and literacy improvement. The aim is to build a strong line of defense for university online ideological security in the intelligent age and promote the organic unity of "tech for good" and educational goals.

KEYWORDS

Generative Artificial Intelligence; Internet Ideological and Political Education; Content Production; Algorithmic Black Box; Value Bias; Safety Governance

1. INTRODUCTION

Currently, Generative Artificial Intelligence (AIGC) technologies, represented by ChatGPT, Sora, and ERNIE Bot, are reshaping the ecology of information production and knowledge dissemination at an unprecedented speed. As a disruptive technology with deep learning, cross-modal generation, and human-like interaction capabilities, Generative AI has not only greatly improved the efficiency and richness of content production but also provided a strong technological engine for the digital transformation of Internet ideological and political education (hereinafter referred to as "Network Ideological and Political Education") in universities. It can generate customized theoretical analyses,

vivid party history stories, and immersive virtual scenes according to students' personalized needs, alleviating to some extent the stubborn ailments of "homogenization, indoctrination, and low interaction" on the supply side of network ideological and political content.

However, as a kind of "Gestell" (Enframing), technology may bring new concealments while revealing the truth [1]. General Secretary Xi Jinping has repeatedly emphasized: "Artificial intelligence has the characteristics of multi-disciplinary integration and high complexity. We must strengthen judgment, coordinate development and security, and control various risks." [2] Unlike traditional retrieval-based Internet tools, the core characteristic of Generative AI lies in "generation" based on probability statistics rather than "search." This generation mechanism relies on the feeding of massive data and parameter adjustments of deep neural networks, and its internal operation is like an impenetrable "Black Box." When this technology is deeply embedded in the content production of network ideological and political education, which has distinct political and ideological attributes, conflicts between technological logic and political logic, as well as algorithmic rationality and value rationality, inevitably surface.

On the one hand, the uninterpretability of algorithms makes the generation process of ideological and political content detached from the complete control of educators, potentially leading to "political errors caused by algorithmic hallucinations." On the other hand, values such as Western-centrism, consumerism, or historical nihilism implied in pre-training data can easily cause "value bias" in generated content through the implicit transmission of algorithmic models, exerting a subtle negative influence on college students who are in a critical period of value formation. In addition, the excessive collection of private data and the technological cession of teacher-student subjectivity also constitute ethical challenges that cannot be ignored.

Therefore, while embracing technological dividends, we must maintain a clear rational consciousness. How to see through the operating mechanism behind the "algorithmic black box"? How to identify and correct potential "value bias"? How to build a governance system that adapts to technological development while ensuring ideological security? This is not only an offensive and defensive battle at the technical level but also a mandatory political question concerning "what kind of people to cultivate, how to cultivate people, and for whom to cultivate people." This paper intends to conduct a theoretical analysis and path construction of the deep-seated problems of Generative AI embedding in university network ideological and political content production from the dual dimensions of ethical risk scrutiny and safety governance.

2. ONTOLOGICAL PERSPECTIVE: THE "ALGORITHMIC BLACK BOX" OF GENERATIVE AI AND THE TRANSPARENCY PARADOX OF IDEOLOGICAL AND POLITICAL EDUCATION

Ideological and political education is a soul engineering project that requires not only "correct results" but also "transparent processes and rigorous logic." However, the technological ontological characteristic of Generative AI—the "Algorithmic Black Box"—has a natural tension with the transparency and certainty required by ideological and political education.

2.1. Technical Causes and Uninterpretability of the "Algorithmic Black Box"

The so-called "algorithmic black box" refers to a complex area between the input and output of an artificial intelligence system that is difficult for human cognition to penetrate. For Generative AI, this black box characteristic mainly stems from the complexity of Deep Learning neural networks.

2.1.1. Hyper-complexity of Parameter Scale

The parameter volume of modern Large Language Models (LLM) has reached the level of hundreds of billions or even trillions (e.g., GPT-4). Massive neuron connection weights are automatically

adjusted through backpropagation algorithms during training, forming an "emergent" ability similar to human intuition. This complex non-linear mapping relationship makes it impossible even for algorithm designers to precisely explain the exact logical path by which the model generates a specific piece of text. In the production of network ideological and political content, this means that when AI generates an elaboration on "the Sinicization of Marxism," we cannot know for sure which specific documents it is based on or what logical reasoning it followed to reach the conclusion.

2.1.2. Stochastic Probability of the Generation Process

Unlike the deterministic logic (If-Then) of traditional programs, Generative AI is essentially an "autoregressive model" based on probability prediction. It predicts the next most likely character based on the preceding context. This probabilistic nature determines that its output has randomness (Temperature). In literary creation, this randomness may bring surprises; but in serious ideological and political education, this "uncertainty" may lead to the blurring or even distortion of political concepts, causing educational content to lose its due rigor and authority.

2.2. Transparency Paradox: The Obscuration of Educational Logic by Technological Logic

Ideological and political education emphasizes "persuading people with reason," requiring educational content to have clear theoretical sources, rigorous logical deduction, and definite value orientation. However, the existence of the algorithmic black box creates a "transparency paradox."

2.2.1. Difficulty in Truth Verification

In traditional teaching, teachers can clearly point out the source of a certain viewpoint (e.g., a certain chapter of *Das Kapital*). However, in AI-generated content, "Machine Hallucinations" often occur, where AI solemnly fabricates non-existent literature, historical facts, or famous quotes. Due to the obscuration of the black box, students and teachers often find it difficult to distinguish truth from falsehood in the first instance, leading to false information spreading in the network ideological and political space under the guise of "intelligence," dissolving the truth power of Marxist theory.

2.2.2. Concealment of Responsible Subjects

The black box structure cuts off the direct chain between behavior and responsibility. When AI generates content containing ideological bias and triggers public opinion issues, should the responsibility be attributed to the algorithm developer, the data provider, the model user (teacher), or the AI itself? This ambiguity of responsibility attribution leaves network ideological and political work facing an ethical dilemma of "no one being responsible," which can easily lead to anomie in educational governance.

3. RISK DIMENSIONS: "VALUE BIAS" AND ITS MANIFESTATIONS IN GENERATIVE AI CONTENT PRODUCTION

If the "algorithmic black box" is a risk in technical form, then "value bias" is the core crisis in content substance. Technology is not value-neutral; every generation by Generative AI is essentially a reproduction and amplification of the values contained in the pre-training data.

3.1. Value Penetration from Data Sources

The value background of Generative AI depends on its training data. Currently, global mainstream large models are mostly developed by Western technology giants, and English data occupies an absolute dominant position in their training corpora, while the proportion of high-quality Chinese corpora, especially Marxist theoretical corpora, is relatively low.

3.1.1. Implicit Implantation of Western Ideology

Since a large amount of training data comes from the Western Internet, the implied Western mainstream ideological spectrums such as "liberalism," "individualism," and "universal values" are internalized by the model as "default settings." When the model processes political concepts involving human rights, democracy, and freedom, it is extremely prone to generating content that conforms to Western narrative logic, while interpretations of concepts such as socialist democracy with Chinese characteristics and whole-process people's democracy may appear biased or vague, causing network ideological and political content to deviate from the mainstream ideological course at the source [3].

3.1.2. Algorithmic Projection of Historical Nihilism

Unofficial histories, historical rumors, and remarks smearing revolutionary heroes existing on the Internet, if mixed into training data without cleaning, may cause AI to unconsciously cite these incorrect historical materials when generating content related to party history and national history. This "banality of evil" based on big data, enhanced by algorithms, may systematically output fragmented views of historical nihilism, seriously eroding college students' historical materialist view of history.

3.2. Instrumental Rationality Alienation of Algorithmic Goals

In addition to data sources, the setting of the algorithmic Objective Function may also lead to value bias.

3.2.1. Catering to Audiences and "Information Cocoons"

To improve user experience and interaction stickiness, algorithms are often designed to "maximize user satisfaction." In network ideological and political scenarios, this may lead AI to tend towards generating content that caters to students' curiosity, emotional needs, or even vulgar tastes, while avoiding serious, profound but potentially dry theoretical interpretations. This substitution of "traffic orientation" for "value orientation" not only leads to the vulgarization of ideological and political content but also intensifies the "information cocoon" effect among students, causing them to stick to existing cognitive biases and refuse to accept guidance from different, especially mainstream, values.

3.2.2. Solidification of Stereotypes and Discrimination

AI models easily capture and amplify existing stereotypes in human society. For example, when generating descriptions regarding occupation, gender, and region, they may unconsciously reveal gender discrimination or regional bias. In ideological and political education content, if AI's description of specific groups (such as impoverished students, ethnic minority students) carries implicit discriminatory colors, it will seriously deviate from the ethical cornerstone of "people-oriented, fairness and justice" in ideological and political education, hurting students' emotions and self-esteem.

4. ETHICAL SCRUTINY: MULTI-DIMENSIONAL CHALLENGES TRIGGERED BY TECHNOLOGICAL EMBEDDING

The intervention of Generative AI in university network ideological and political education not only triggers risks at the content level but also poses challenges to teacher-student subjectivity, privacy security, and educational ecology in deep ethical relationships.

4.1. Dissolution of Subjectivity: From "Human-Machine Collaboration" to "Technological Dependence"

Marx pointed out: "Machinery adapts itself to the weakness of the human being in order to turn the weak human being into a machine." [4] Under the coercion of intelligent technology, the crisis of subjectivity in ideological and political education is becoming increasingly prominent.

4.1.1. Degeneration of Teachers' Subjectivity in "Teaching"

Some ideological and political teachers may develop technological dependence due to the high efficiency and convenience of AI, entrusting core work such as lesson plan writing, courseware production, and Q&A entirely to AI, while they themselves degenerate into "operators" clicking the "generate" button. This trend of "de-skilling" causes teachers to gradually lose the motivation for theoretical study and the ability for independent thinking, leading to the loss of the inner soul of "educating oneself before educating others" in ideological and political education, reducing the educational process to cold data interaction.

4.1.2. Suspension of Students' Subjectivity in "Learning"

For students, Generative AI may become a "cheat sheet" for obtaining standard answers. When facing ideological and political homework or theoretical thinking, if students are accustomed to directly asking AI for answers and abandoning deep reading and reflection, the cultivation of their critical thinking and value judgment abilities will be blocked. This phenomenon of "thinking outsourcing" makes students porters of knowledge rather than constructors of meaning, seriously violating the original intention of fostering character and civic virtue in ideological and political education.

4.2. Erosion of Privacy Boundaries: Data Mining and Ethical Bottom Lines

Precise network ideological and political education often relies on the analysis of student big data, and the interactive nature of Generative AI exacerbates the risk of privacy leakage.

4.2.1. Imperceptible Collection of Sensitive Data

In the process of conversing with AI, students may unconsciously reveal highly sensitive information such as personal political leanings, psychological confusion, and family conditions. If this data is indiscriminately collected by the model backend and used for iterative training, or even abused by commercial institutions, it will seriously infringe upon students' privacy rights.

4.2.2. Ethical Overstepping of "Data Profiling"

AI conducts precise profiling and prediction of students based on data. While helpful for individualized instruction, it may also evolve into a kind of "data panopticon." Students become "naked" in front of algorithms, with their ideological dynamics quantified, predicted, and intervened. This excessive data surveillance may trigger resentment and resistance among students, destroying the trust foundation required for ideological and political education, and even triggering an ethical crisis of "algorithmic control."

4.3. Falsity of Emotional Interaction: Indifference Behind Anthropomorphism

Ideological and political education is "emotional education," requiring the investment of genuine feelings. Although Generative AI can simulate human tones for "anthropomorphic" communication and even provide "emotional value," its essence remains cold inorganic matter.

4.3.1. Risk of Emotional Deception

When students confide their troubles to AI and receive seemingly warm comfort, they may have the illusion of establishing an emotional connection with AI. However, this connection is one-way and illusory. AI's emotional feedback is merely the "optimal solution" calculated by algorithms, not care

from the heart. This "false emotional presence" may lead to students' withdrawal in real interpersonal communication, and even a deeper sense of nothingness and deception after discovering the truth of AI's "heartlessness."

4.3.2. Absence of Life Education

AI cannot experience the pain, joy, and sublimity of life, so it cannot truly understand and impart deep ideological and political topics about the meaning of life and ideals and beliefs. Relying entirely on AI for network ideological and political education will lead to the "de-life-ization" of educational content, turning ideological and political education into a logical game lacking life temperature.

5. GOVERNANCE PATHS: BUILDING A SECURITY BARRIER FOR NETWORK IDEOLOGICAL AND POLITICAL EDUCATION IN THE INTELLIGENT AGE

Facing the risks of algorithmic black boxes and value bias brought by Generative AI, we should not stop eating for fear of choking but should adopt a strategy of "equal emphasis on defense and governance, and co-governance of technology and institutions" to build a comprehensive safety governance system.

5.1. Technical Governance: From "Black Box" to "Trustworthy"

Improving algorithmic transparency and value alignment capabilities through technical means is the foundation of governance.

5.1.1. Constructing Vertical Large Models (Vertical LLM)

Universities should collaborate with research institutions to train or fine-tune vertical domain large models for ideological and political education based on high-quality data such as the Marxist theoretical system, CPC history, and national and social conditions. Establish a strict "whitelist system for pre-training corpora" to ensure the political correctness and value purity of input data, cutting off the penetration of incorrect values from the source [5].

5.1.2. Strengthening Reinforcement Learning from Human Feedback (RLHF)

Establish a "Human-in-the-loop" optimization mechanism. Organize high-level ideological and political teacher teams as "labelers" and "evaluators" to score and correct the values of AI-generated content. Translate human political judgments and moral norms into reward signals fed back to the model, and through continuous iterative training, achieve deep alignment between AI's value orientation and socialist core values.

5.1.3. Developing Explainable Artificial Intelligence (XAI) Technology

Promote research on algorithmic explainability and develop plug-in tools capable of displaying generation logic and tracing cited literature. In network ideological and political applications, mandatorily require AI to provide a "chain of evidence" for the generation of key political viewpoints, allowing teachers and students to know the how and the why, breaking the mystery and misleading nature of the algorithmic black box.

5.2. Subject Reshaping: Establishing Ethical Norms for "Human-Machine Collaboration"

Reshaping the subject status of humans (teachers and students) in the intelligent technological environment is the core of governance.

5.2.1. Establishing a "Teacher-Led" Review and Gatekeeping Mechanism

Clearly define the "auxiliary" status of Generative AI. In the production and release process of network ideological and political content, establish a strict three-level review system of "AI draft—Teacher review—Political gatekeeping." It is stipulated that content involving major political theoretical issues must be written by humans or undergo substantive human modification, and the final right of release must never be ceded to algorithms. Teachers must shoulder the responsibility of "value gatekeepers" and bear the final responsibility for the political orientation of AI-generated content [6].

5.2.2. Enhancing Teachers' and Students' Algorithmic Literacy and Digital Ethics

Incorporate algorithmic literacy into the ideological and political education system. On the one hand, train teachers to master "Prompt Engineering" skills for conversing with AI to improve their ability to harness technology; on the other hand, educate students to correctly recognize the limitations and biases of AI, cultivate critical thinking, and learn to scrutinize and screen AI-generated information, thereby achieving "human-machine co-intelligence" rather than "human-machine conflict."

5.3. Institutional Regulation: Perfecting a Full-Cycle Risk Prevention and Control System

Drawing red lines and bottom lines for technology application through institutional construction is the guarantee of governance.

5.3.1. Formulating Ethical Guidelines for Generative AI Educational Application

Universities should issue special Administrative Measures for the Application of Generative AI in Network Ideological and Political Education, clarifying boundaries for data collection, standards for content generation, and responsibilities for violations. Establish "algorithmic filing" and "ethical review" systems to conduct security assessments of AI models introduced to campus, preventing models with malicious code or value backdoors from entering the campus network.

5.3.2. Establishing Dynamic Risk Monitoring and Circuit Breaker Mechanisms

Use technical means to establish a public opinion monitoring system for network ideological and political education, focusing on monitoring the dissemination effects and public opinion feedback of AI-generated content. Once AI-generated content is found to trigger ideological risks or ethical controversies, immediately trigger the "circuit breaker mechanism," suspend related services, and conduct manual intervention and model repair to prevent the spread of risks [7].

5.3.3. Building a Firewall for Data Privacy Protection

Strictly implement the Data Security Law and the Personal Information Protection Law. Promote "localized deployment" of data and "available but invisible" technologies (such as privacy computing). Ensure that students' ideological dynamic data circulates in a closed loop within the school, strictly prohibiting unauthorized transmission of sensitive data to third-party commercial models, effectively safeguarding national security and student rights [8].

5.4. Ecological Optimization: Creating a Healthy and Rational Intelligent Education Environment

5.4.1. Promoting Algorithmic Weighting of Mainstream Values

Actively introduce "mainstream value weights" in the recommendation algorithms of campus network ideological and political platforms. Instead of relying solely on click-through rates and completion rates as the basis for recommendation, significantly increase the recommendation weight of content with ideological, theoretical, and educational value. Through algorithmic intervention, break the "information cocoon" and make positive energy content the "big traffic" in cyberspace [9].

5.4.2. Returning to the Essence of Education in Interpersonal Interaction

Advocate a blended education model of "online empowerment, offline foundation." Use AI to solve efficiency problems in knowledge transfer, thereby freeing up more time for teachers to return offline to conduct face-to-face deep counseling, heart-to-heart talks, and practical teaching. Use the teacher's personal charisma and sincere emotions to make up for the lack of temperature in AI technology, building a new educational ecology where technology and humanities shine together.

6. CONCLUSION

The wave of Generative AI is surging, opening the door to the future for university network ideological and political education while also casting the shadow of algorithmic black boxes and value bias. This is a profound game involving technology and values, efficiency and ethics. Scrutinizing ethical risks is not to deny the value of technology, but to harness technology more soberly and safely.

The digital transformation of university network ideological and political education must adhere to the ethical purport of "tech for good" and always take fostering character and civic virtue as the starting point and goal of technology application. By constructing a safety governance system that is "technologically trustworthy, subjectively conscious, institutionally complete, and ecologically healthy," we are fully capable of taming the "algorithmic black box," correcting "value bias," and making Generative AI a "golden microphone" for disseminating Marxist truth rather than an amplifier for breeding erroneous trends of thought. In the new panorama of human-machine collaboration, we will continuously create a new situation for the high-quality development of university network ideological and political education and cultivate newcomers of the era who possess both digital literacy and firm beliefs.

ACKNOWLEDGEMENTS

- (1) 2025 GDUF School-Level Project on Marxist Theory and Ideological-Political Education “Mechanisms and Pathways for Generative-AI-Enabled Production of Online Ideological-Political Content in Universities” (Project No. 25XJ23)
- (2) 2025 GDUF School-Level Project on Marxist Theory and Ideological-Political Education “Mechanisms and Pathways for Cultivating University Students’ Patriotism through Digital Narratives of Red Finance in the New Era” (Project No. 25XJ06)

REFERENCES

- [1] Chen Xiaoping. Philosophical Reflection and Governance Path of Artificial Intelligence Ethical Risks [J]. Philosophical Trends, 2023(08): 89-96.
- [2] Xi Jinping. Adhering to the Overall National Security Concept and Blazing a Path of National Security with Chinese Characteristics [N]. People's Daily, 2014-04-16(01).
- [3] Sun Hanqi. Ethical Risks and Governance of Generative Artificial Intelligence Empowering Ideological and Political Education [J]. Journal of Ideological and Theoretical Education, 2025(01): 112-118.
- [4] Marx, Engels. Complete Works of Marx and Engels (Vol. 42) [M]. Beijing: People's Publishing House, 1979: 95.
- [5] Yang Yunxia. Adherence to Subjectivity and Technological Empowerment of Ideological and Political Education in the Intelligent Age [J]. Studies in Ideological Education, 2025(03): 22-28.
- [6] Wang Shao. Technical Route, Security Risks and Prevention of ChatGPT Intervention in Ideological and Political Education [J]. Journal of Shenzhen University (Humanities & Social Sciences), 2023, 40(02): 153-160.
- [7] Feng Lin, Ni Guoliang. Digital Transformation of Ideological and Political Education Based on Generative Artificial Intelligence [J]. Studies in Ideological Education, 2024(02): 46-53.

- [8] Liu Xiaofeng, Zhang Jinlin. Three-Dimensional Exploration of Generative Artificial Intelligence Impacting University Ideological and Political Education [J]. Journal of National Academy of Education Administration, 2023(12): 66-75.
- [9] Deng Huan. Research on the Path of Precise Supply of University Internet Ideological and Political Education from the Perspective of Generative Artificial Intelligence [J]. China Higher Education, 2025(02): 55-60.