ISSN Print: 2707-0638 Online: 2707-0646
New Exploration of Ideology and Politics − 2026, 8(2)
https://doi.org/10.35534/neip.0802027

Multiple Dimensions of Ideological Risks in Generative Large Language Models and Countermeasures

Shengyang Luo

Changchun University of Technology, Changchun, China

Abstract: Based on their revolutionary capabilities in information comprehension and text generation, generative large language models (LLMs) have been widely applied in production and daily life. The deep integration of their technological applications with social structures not only carries ideological attributes but also poses the risk of multi-dimensional proliferation regarding political direction, value orientation, and public opinion guidance. Driven by the practical need to address the ideological risks of generative LLMs, it is urgent to uncover their ideological attributes—from macro-theoretical interpretation to micro-application observation—and specifically analyze the causes of these risks. Furthermore, effective approaches to countering these ideological risks should be explored across three dimensions: ideological awareness, institutional safeguards, and technological measures.
Keywords: Generative Large Language Models; Ideology; Potential Risks; Governance Strategies

Currently, generative large language models (LLMs) are reshaping traditional modes of production and daily life with their unprecedented knowledge processing capabilities. The boundaries of social interaction have been greatly extended, traditional constraints of time and space have been removed, and the entire social sphere is ushering in broad and profound transformations. First, the illusion of 'technological neutrality' disperses mainstream ideology. The existence of algorithms has transformed into coherent text and structured information. Consequently, the pre-existing value-laden and non-neutral implications embedded in training data—such as stances on social issues, the output of cultural values, and gender or racial biases—are endowed with a false shell of scientificity. This enables them to accomplish an implicit dissolution of mainstream ideology within seemingly objective technological operations. Therefore, clarifying the ideological attributes of LLMs and analyzing their inherent ideological risk mapping (Figure 1), thereby providing a perspective on the ideological risks they may induce, is crucial for understanding the essential attributes of LLMs and circumventing these ideological risks.

Funding: Jilin Province Higher Education Scientific Research Project "Research on the Effectiveness of AI Technology Empowering Ideological and Political Education in Universities" (JGJX25D0257); Jilin Educational Science Planning Project "Research on AI-Empowered Digital Teaching Reform Strategies in Higher Education of Jilin Province" (GH25640).
Author Introduction: Shengyang Luo is a Master's candidate at the School of Marxism, Changchun University of Technology. His research focuses on artificial intelligence and Marx's view on machinery.
Article Citation: Luo, S. Y. (2026). Multiple Dimensions of Ideological Risks in Generative Large Language Models and Countermeasures. *New Exploration of Ideology and Politics, 8*(2), 309–320.

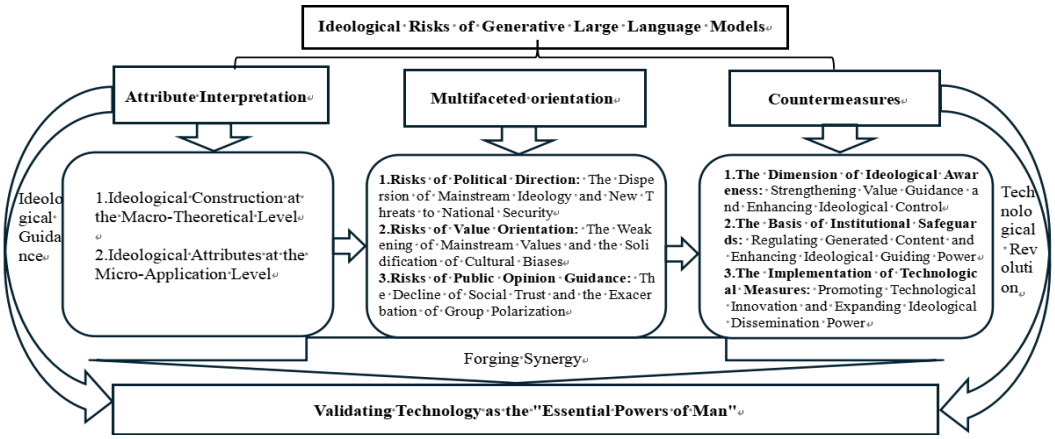

Figure 1 Conceptual Framework for Countering the Ideological Risks of Generative Large Language Models

1 Interpretation of the Ideological Attributes of Generative Large Language Models

Generative large language models are intelligent technologies based on deep learning, particularly advanced neural network architectures such as the Transformer. Pre-trained on massive textual data, they master the statistical laws and semantic knowledge of language, thereby acquiring the capabilities to understand, generate, and transform natural language (Miao, Wang, & Yang et al., 2024). To elucidate why ideological attributes refer to the value positions, political orientations, and cultural biases embedded in data, algorithms, and output content. Generative large language models are inherently imbued with an ideological undertone, and the primary task is to clarify the relationship of science and technology to ideology at a macro-theoretical level. By examining the rationality and one-sidedness of the "Neutrality Theory," the "Opposition Theory," and the "Identity Theory," a scientific interpretation of the dialectical relationship of science and technology to ideology is unfolded. Responding to the realistic connection between generative large language models—as technological products—and ideology from the standpoint of historical materialism is the prerequisite for investigating their ideological attributes, and governance approaches from an applied dimension.

1.1 The Ideological Construction of Generative Large Language Models at the Macro-Theoretical Level

In the ideological domain, academic discussions regarding the relationship of science and technology to ideology remain contentious. A basic consensus has yet to be reached, and theoretical divergences among the "Neutrality Theory," the "Opposition Theory," and the "Identity Theory" have long persisted. The "Neutrality Theory" argues that science is an enterprise pursuing pure truth, its essence is value-free, and objectivity is the lifeblood of science, which is entirely divorced from value, subjectivity, and subjective factors. This theory, which forcibly separates science from value and ideology, ostensibly defends the "objectivity" of science. In reality, however, it severs the connection between subject and object in scientific practice, dissipates value factors with so-called objectivity, and degrades science to a mere technological tool, ironically providing a cloak of legitimacy for the capitalist abuse of science. The "Opposition Theory," represented by the French structuralist Marxist Louis Althusser, contends that a fundamental

dichotomy separates science and technology from ideology. This absolutized dichotomy oversimplifies the complexity of the issue, constituting an isolated and one-sided understanding of both forms. It not only denies their intricately intertwined relationship in real society and history but also fails to recognize the unified relationship between truth and value inherent in science and technology, as well as in ideology, leaving irreconcilable tensions within the theory itself. The "Identity Theory," represented by the Frankfurt School, affirms the thesis that science and technology are also a form of ideology. Proceeding from the context of advanced industrial society, it reveals a new essential connection between technology and ideology. Prior to advanced industrial society, technology was considered value-neutral; even if acknowledged to be value-laden, its relatively weak power precluded it from becoming a form of domination. However, in advanced industrial society, due to the hegemonic status of science and technology, the domination of man by man partially gives way to an objective, reified mechanism—namely, the manipulation of technological rationality. Thus, technological rationality participates in the construction of ideology (Liu, 2010). This theoretical reflection was proposed by Max Horkheimer, further developed by Herbert Marcuse, and brought to theoretical and systematic culmination by Jürgen Habermas. Undeniably, the Frankfurt School soberly recognized the profound impact of science and technology on contemporary capitalist society and revealed the fact that science and technology, as a form of power, provide apologetics for the bourgeoisie. However, the "Identity Theory" attributes the fundamental contradictions, cyclical crises, and the roots of antagonism, conflict, and disaster in capitalism to technology itself. By substituting the critique of technological rationality for the critique of political economy, it fails to touch upon the economic and institutional roots behind social facts, thereby falling into the old rut of merely blaming science and technology, ultimately moving towards the opposite of its critical original intention.

Marx's reflection on the relationship of science and technology to ideology proceeds from a dialectical perspective. Distinct from the cognitive limitations of the "Neutrality Theory," "Opposition Theory," and "Identity Theory," Marxism holds that science and technology are related to ideology but not identical to it, distinct yet not oppositional. Specifically, the first aspect is distinguishing science and technology from ideology. Firstly, the objects they reflect are different. Science and technology reflect the objective laws of nature. In the Preface to *A Contribution to the Critique of Political Economy*, Marx emphasized that natural science reveals the laws of material transformation that "can be determined with the precision of natural science," (Marx & Engels, 2009a: 592) studying objective natural processes independent of human will. Ideology, on the other hand, reflects the interest relations of specific social classes. In essence, it is a "theoretical system of values" and a value judgment reflecting different interest relations, with its core lying in providing legitimizing apologetics for specific class rule. Secondly, their inherent attributes differ. Science and technology possess objective truthfulness; their truthfulness lies in conforming to objective reality, being verifiable through practice, and possessing a universal applicability that transcends class. Ideology possesses a class-value nature; it is the theoretical expression of specific class interests, concealing substantive interest particularity with a formally "illusory" universality of interests, thus possessing distinct class attributes. Finally, their functions differ. Science and technology are a historically revolutionary, driving power (Marx & Engels, 2009b: 602), directly promoting production development and social progress. Ideology, as part of the superstructure, functions to maintain or subvert specific interest paradigms. By portraying "special interests as general interests," it provides a rational defense for class domination. The second aspect is the connection of science and technology with ideology. On the one hand, science and technology are inherently revolutionary. Technological renewal profoundly transforms social production and lifestyles while simultaneously destroying past backward ideologies, realizing ideological renewal and development. On the

other hand, politics, law, and morality, as manifestations of ideology, possess relative independence and actively react upon the development of science and technology. Consequently, they accelerate or retard the progress and practical application of science and technology, governing their social nature and developmental trajectory.

1.2 The Manifestation of Ideological Attributes of Generative Large Language Models at the Micro-Application Level

Confirming the ideological attributes of generative large language models from the perspective of historical materialism is a theoretical interpretation at the macro level. We may focus on the micro-application dimension to examine the ideological attributes of generative large language models—as the objectification of science and technology—at the level of practical application. As a technological product, its essence is not a purely computational tool characterized by technological justice and objective neutrality. Instead, it may embed ideological attributes within operational links such as technological architecture and application fields, presenting a realistic orientation where science and technology carry ideology, thereby infiltrating and influencing people's thoughts and behaviors.

The reasons are as follows. First, during the process of data collection, analysis, and integration, due to the inherent limitations of the computational data itself, every step from data acquisition to model deployment may introduce or exacerbate social, cultural, and technological biases. Existing datasets and corpora are mostly "imported" data materials, which inherently carry the ideological biases brought about by their original environments (Wu, 2024). When biased data samples enter the machine learning system and go through numerous steps relying on human annotation—such as data cleaning, pre-processing, and content filtering—their language output is easily influenced by the original data biases and the subjective judgments of the annotators. Second, as the model utilizes algorithmic architecture as its analytical mode, its internal value embedding tends to be concealed. In the process from LLMs receiving instructions to generating responses, the stage from instruction input to pre-processing and understanding may misinterpret the original cultural background; the stage from understanding to decoding and sampling may be subject to the transmission and shaping of the specific values of the annotators; and the process from sampling strategy to the final text output may embed subtle ideological guidance through the adjustment of the expression framework. Each link provides the possibility for ideological infiltration. Meanwhile, the internal complexity of the model largely isolates external supervision, resulting in a lack of interpretability and transparency in its generated content, which cannot be effectively traced. Third, the content generation of the model is constrained by certain social relations. In the early stages of their development and evolution, science and technology often carry the beautiful vision of promoting the common progress of all humanity. However, the participation of technological media has not changed the fundamental relations of production. This idealized blueprint is still subject to real capitalist relations of production in reality, casting a shadow over its initial commitment to universal benefit. Entering the AIGC (Artificial Intelligence Generated Content) era, the exploitation of capital logic has become more concealed, making artificial intelligence inherently contain the risk of capitalist ideological domination while promoting the progress of human society. Marx pointed out that "It takes time and experience for the workers to learn to distinguish between machinery and the capitalist use of machinery, and to direct their attacks not against the material instruments of production but against the mode of their use." (Marx & Engels, 2009c: 493) From events ranging from the early Luddite movement to the recent Hollywood strikes, Marx's insights from over a century ago remain far from outdated. The numerous risks brought about by current generative large language models are closely related to capital-driven and profit-oriented application logic in practice. Driven by

profit-seeking motives, the prudential regulation of technological application risks is often shelved, leading to symptoms of ideological risks across various dimensions such as politics, values, and public opinion.

2 Multiple Dimensions of Ideological Risks in Generative Large Language Models

With the widespread application of intelligent technologies, generative large language models, represented by DeepSeek and ChatGPT, are entering millions of households through their technological convenience, constructing ideological agents with code as their bones and data as their flesh and blood. Relying on the application of deep learning, attention mechanisms, and neural networks, generative large language model technology has greatly enhanced information processing and knowledge integration capabilities. However, its misapplication has also spawned uncertain variables affecting ideological security, presenting undeniable ideological security risks within the three-dimensional framework of political direction, value orientation, and public opinion guidance.

2.1 Risks of Political Direction: The Dispersion of Mainstream Ideology and New Threats to National Security

Marx pointed out: "If considered from an ideal standpoint, the dissolution of a given ideology is sufficient to cause the downfall of an entire epoch." (Marx & Engels, 2009d: 170) This reveals the fundamental role of conceptual systems in the survival of an era. Viewed from this perspective, the primary risk of generative large language models points directly to the fundamental issue of political direction. Under the new paradigm where technological power is deeply involved in content production, the risk of technological potential energy transforming into political influence arises, subjecting mainstream ideology to the possibility of being diluted, obscured, or even displaced.

First, the illusion of "technological neutrality" erodes mainstream ideology. The existence of technology itself "inherently" possesses a certain value inclination and is not politically neutral; it inherently has its own preferences in national governance (Shang & Liu, 2025). As mentioned previously, defining science and technology as objective entities obviously provides an exonerating rhetoric for the infiltration of heterogeneous values. Under the logic of content production dominated by generative large language models, data is systematically processed by algorithms and transformed into coherent text and structured information. Consequently, the pre-existing non-neutral realistic expressions underlying the data—such as stances on social issues, the output of cultural values, and gender or racial biases—are endowed with a false shell of scientificity. This enables them to accomplish an implicit dissolution of mainstream ideology within seemingly objective technological operations. And when content generation deviates from the guidance of mainstream ideology and submits to so-called "value-free" science and technology, social consciousness loses its effective pathway to react upon social existence through content, and its value-guiding power, cohesive power, and inspiring power are subsequently weakened.

Second, LLM-facilitated disinformation and deepfake-like content. Deepfakes plunge the public into a cognitive dilemma where truth and falsehood are difficult to distinguish, through methods such as resetting reality and fabricating facts (Li & Li, 2025). The content generation of large language models is not a faithful reproduction of reality; rather, it relies on structural probabilities implicit in the training data, not on true top-level design or deep reasoning (Zhang, 2025). This type of narrative fabrication, masked by technology, has already become the most dangerous transmission

path for historical nihilism, causing social facts to be replaced by customizable digital products. Specifically, when dealing with controversial issues, the probability-generating mechanism dominated by algorithms tends to replace strict statements of fact with seemingly self-consistent narratives. It weakens traditional media's adherence to objective truth and, leveraging fragmented transmission methods, creates a post-truth digital context where social reality gives way to subjective imagination. Once controlled by external forces, such a subversive tool of political expression can leverage its low cost and speed to mass-produce inflammatory simulation information, instigate social unrest at extremely low costs, and thereby pose a severe challenge to national political security.

2.2 Risks of Value Orientation: The Weakening of Mainstream Values and the Solidification of Cultural Biases

The value orientation risk of generative large language models focuses on the value security embedded within the model's own data and algorithmic framework. Generative large language models are an aggregate composed of technological elements such as data and algorithmic frameworks. The value orientation of their text generation is inevitably correlated with the selected corpora and algorithmic applications, planting hidden dangers for the biases in the values they generate. Among these, the technology-associated problems of "data bias" and "algorithmic black boxes" harbor the latent risks of diluting mainstream values and leading to the marginalization of cultural subjectivity.

First, "data bias" weakens the authority of mainstream values. Large-scale databases have integrated human knowledge on an unprecedented scale and density; however, in this process, they quietly transform statements that originally belonged to specific cultural contexts, historical narratives, and value systems into "established facts", enabling specific ideologies to be reproduced within the algorithmic logic. The distribution of corpora in the database determines the model's generation probability. For cultural expressions with higher frequencies and more unified narratives, the model tends to fit them with greater weight, thereby repeating or amplifying the value presuppositions they carry when generating responses. This content encoding, accomplished through statistical mechanisms, naturalizes culturally biased narratives into "more rational expressions" within the probability space. Conversely, marginalized cultures, peripheral narratives, and non-mainstream values are diluted or even dissolved during algorithmic compression. Consequently, a dislocation forms between the superficial language generation of the model and its deep value structure, causing the cultural content outputted by data symbols to inherently carry the value intentions of specific ideologies, continuously challenging and deconstructing the systematic and structured construction of mainstream ideological content.

Second, the "technological black box" solidifies cultural biases. The "technological black box" refers to the non-linear, opaque process by which algorithmic systems transform data inputs into information outputs (Jiang, 2024). The complexity of the generation mechanisms of large language models results in a lack of interpretability in their text and image data generation processes, making it difficult for people to understand the logic of information generation. Meanwhile, discourse outputs that undergo black-box generation—such as tendencies of historical nihilism and cultural biases appearing in generated texts—are difficult to identify and trace due to the invisible state of steps like corpus selection, narrative elaboration, and emotional processing. This provides ample opportunities for the infiltration of erroneous ideological trends and harmful ideologies. The ambiguity of value definition and the uncertainty of responsibility attribution in generative large language models leave generated content suspended outside of regulatory

and legal frameworks, effectively forming a regulatory gray zone for content generation. In the absence of text interpretability and algorithmic visibility mechanisms, accompanied by technological iteration and the enhancement of machine learning autonomy, the algorithmic black box effect is continuously amplified, gradually pushing value expression into a grey area.

2.3 Risks of Public Opinion Guidance: The Decline of Social Trust and the Exacerbation of Group Polarization

In addition to the security risk patterns in the political and value dimensions, generative large language models also pose risks in public opinion guidance, manifested as the erosion of social trust and the acceleration of group polarization. This public opinion guidance issue triggered by technology is reflected at the model level as a dual superposition of the "illusion of customization" and "emotional arousal," spawning systemic challenges to public rationality and group cognition.

First, the "illusion of customization" leads to a decline in social trust. Through deep learning and semantic refinement of massive data, LLMs rapidly generate narrative content that meets user needs, creating an "illusion of customization" where users feel they completely control content generation. This makes it highly easy to view the algorithm-generated results as a direct extension of one's own will, rather than the product of a complex computational process (Xiang, 2025). Such customized narratives can not only amplify certain information but also imperceptibly guide the cognitive direction of the public. By continuously iterating and enhancing technological functions to strengthen the connection with users, and gradually building users' instrumental trust and reliance, it is easy for users to fall into the rut of technological dependency. Consequently, social credibility is increasingly weakened in narratives where the virtual and the real intertwine. It becomes difficult for the public to discern authenticity in technological narratives. This not only leads to the anomie of social ethical relations but also, once suspicions regarding the trustworthiness of technology are formed, every individual or nation will be unwilling to make further trusting or cooperative moves beyond their own known information and rational evidence. Thus, falling into the "Tacitus Trap" (Zhu, 2026) in the digital era, triggering a widespread and profound crisis of trust.

Second, "emotional arousal" causes the tearing of group consensus. The information sources of traditional media are authoritative, and their transmission channels undergo multifaceted scrutiny and multi-layered review. In contrast, generative systems, due to the bias of training data and the absence of "gatekeepers," easily produce pseudo-self-consistent, inflammatory, and one-sided information, thereby covertly transmitting and infiltrating homogenized attitudes and values in their interactions with humans." (Wang, 2025) Catalyzed by emotional resonance, such information spreads rapidly, reinforcing internal group identity while exacerbating the rejection of external viewpoints. This encloses users within machine-generated, reinforces their perception of inherent biases, and thereby accelerates emotional antagonism between groups. This phenomenon, to a certain extent, weakens the dialectics and diversity of public discourse, tending to narrow the space for public discussion and making dialogue between different groups increasingly difficult. When individuals are addicted to the emotionally resonant strata woven by algorithms, the space for rational dialogue is continuously compressed, and the achievement of consensus increasingly relies on emotional side-taking rather than factual discrimination. As Neil Postman warned: "Technopoly will eventually reduce culture to a mere appendage of entertainment." Driven by such emotions, social consensus gradually gives way to the efficiency of communication and the intensity of emotion, allowing ideological risks to breed imperceptibly.

3 Countermeasures Against the Ideological Risks of Generative Large Language Models

With the advent of the intelligent digital era, the overt trend of ideological confrontation is slowly unfolding within the technological space. Advancing ideological work as a systematic project is the top priority of our current endeavors. This requires us to holistically promote ideological governance within a three-dimensional framework of awareness, institutions, and technology, with the expectation of achieving a balance between the technological advancement of generative large language models and risk management.

3.1 The Dimension of Ideological Awareness: Strengthening Value Guidance and Enhancing Ideological Control

Mainstream ideology serves as the ideological guide for technological development. Strengthening the hegemony of mainstream ideology is a necessary measure for technological progress and the shaping of subjective values. It is imperative to fundamentally enhance the ideological control of mainstream ideology over LLM technologies.

First, strengthening the leadership of mainstream socialist ideology over digital technology platforms. Faced with the containment and blockade by Western forces in the ideological domain, establishing the leadership of socialist ideology has become a primary task. The ideological confrontation on digital technology platforms is, in essence, a conflict of viewpoints and cultures between different civilizations behind the technology. Therefore, it is necessary to guide the value biases of ideology to prevent potential security risks in the conceptual dimension. Specifically, we need to select data sources of mainstream ideology, promote the embedding of normative socialist values into the system design of machine learning, and use mainstream value orientation to guide non-institutionalized digital narratives. In a clamorous and pluralistic value ecosystem, unifying people's minds with a unitary core to achieve value alignment between machine intelligence and mainstream ideology is the methodological principle for further enhancing ideological control and strengthening value guidance.

Second, promoting the guidance and shaping of technology R&D subjects by mainstream values. The foundational cognition of professional and technical personnel has a direct impact on the entire operational process of generative large language models. There is an urgent need to strengthen the guiding role of mainstream values for R&D personnel, enabling them to actively identify and avoid risks that may exacerbate social biases or infringe upon digital rights with their firm political, moral, and value stances. Through the value shaping of technological subjects, a contingent of technical talents possessing both moral integrity and professional competence should be cultivated. This integrates the reserve of knowledge and skills with value consensus, ethical cognition, and political identification, thereby holistically advancing the enhancement of AI literacy among technology developers. This means that the R&D and creation of technical professions are not merely about solving technical problems; they also bear the social responsibility of consolidating and expanding the cohesion and inspiring power of mainstream social values.

Third, promoting the embedding of mainstream values into the operational process of generative large language models. Throughout the operational process of generative large language models, the embedding of mainstream values must permeate every link, including model training, content generation, and interactive feedback. Specifically, in the model training stage, priority should be given to high-quality corpora reflecting socialist core values. Value alignment should be achieved through Retrieval-Augmented Generation (RAG) and fine-grained annotation technologies, rectifying

the value biases of the model's corpora from the source. In the algorithm design stage, the profit-first logic of capital valorization should be abandoned, highlighting the core stance of the algorithmic logic that takes the people as the principal body. In the content output stage, measures should be taken to enhance the weight of content conforming to mainstream values, strengthen the explanatory power, broad-spectrum reach, and influence of mainstream value concepts, and promote generative large language models to become a technological increment for mainstream value identification.

3.2 The Basis of Institutional Safeguards: Regulating Generated Content and Enhancing Ideological Guiding Power

Xi Jinping (2018) pointed out: "We must strengthen the assessment and prevention of potential risks in the development of artificial intelligence, safeguard the interests of the people and national security, and ensure that artificial intelligence is safe, reliable, and controllable." To promote the healthy development of LLMs and ensure data security and stability in the ideological domain within their application scenarios, it is necessary to strengthen the construction of the risk prevention and control system and synergistically advance the dynamic articulation of institutions and laws.

First, strengthening the value guidance of the institutional system is the prerequisite for regulating technological development. The *Resolution of the Central Committee of the Communist Party of China on Further Deepening Reform Comprehensively to Advance Chinese Modernization*, adopted at the Third Plenary Session of the 20th Central Committee of the Communist Party of China, proposes to " improve the mechanisms for developing and managing generative artificial intelligence" and "strengthen the cybersecurity system and institute oversight systems to ensure the safety of artificial intelligence", providing strategic guidance for the developmental direction and governance framework of generative large language models (Chinadaily.com.cn., 2024). Based on the increasingly apparent ideological attributes in the development of LLMs, strengthening value norms within the institutional system is essential to better promote AI governance. Accordingly, China promptly issued the *Global AI Governance Initiative*, proposing that "We should uphold a people-centered approach in developing AI, with the goal of increasing the wellbeing of humanity and on the premise of ensuring social security and respecting the rights and interests of humanity, so that AI always develops in a way that is beneficial to human civilization."(Cyberspace Administration of China, 2023) The construction of the institutional system aims to better regulate the application scenarios of AI. In its application, we must always adhere to the value guidance of "people-centeredness" and "AI for good," incorporate mainstream ideology into the stipulations of top-level design, and make AI a "booster" for scientifically increasing human well-being rather than a "stumbling block."

Second, constructing AI laws that consolidate fundamentals, stabilize expectations, and benefit the long term is the foundation of technological progress. At the current stage, China has formed a foundational institutional framework encompassing the *Provisions on the Administration of Algorithm-Generated Recommendations for Internet Information Services,* the *Measures for the Administration of Internet Information Services,* the *Ethical Norms for the New Generation Artificial Intelligence*, and the *Interim Measures for the Management of Generative Artificial Intelligence Services.* A multi-level legal and regulatory system in the field of generative large language models has been constructed, making systematic stipulations on normative requirements in links such as technological R&D and service provision. In the long run, the current legal system has, to a certain extent, regulated the political, value, and public opinion risks of generative large language models and provided a reference framework for compliance practices in related fields. However, the overall system still exhibits obvious gaps and lags in terms of the unity of governance subjects, the

forward-looking nature of governance norms, and the enforceability of governance rules (Li & Li, 2025), making it difficult to cover and effectively bridge the complex risks brought about by technological iteration. Accordingly, it is necessary to firmly root in data, algorithms, and computing infrastructure to further perfect the institutional norms for generative large language models, ensuring their steady operation on the track of the rule of law, thereby firmly holding the bottom line of ideological security.

Third, advancing pluralistic social governance and supervision is the guarantee for technological operation. The practice of relying solely on top-level macro-policies and meso-level laws and regulations has obvious limitations. There is an urgent need to gather the forces of all sectors of society to make pluralistic governance a positive support for the steady operation of technology. Therefore, a pluralistic co-governance social system should be formed, integrating the leadership of Party organizations, the public opinion supervision of mainstream media, the fulfillment of responsibilities by industry organizations, and the self-regulation of social organizations (Zhou & Jin, 2024). The principled nature of top-level design should be organically combined with the flexibility of pluralistic governance to clarify the boundaries of power and responsibility of various functional subjects, and to improve the synergistic linkage mechanism of laws and regulations under a multidimensional governance framework. This will effectively prevent the ideological control and risk spillovers that LLMs may bring. Concurrently, intelligent technology should be deeply embedded in regulatory practices to further alleviate information asymmetry in supervision with powerful data processing capabilities, thereby supplementing the enhancement of regulatory efficacy and precision.

3.3 The Implementation of Technological Measures: Promoting Technological Innovation and Expanding Ideological Dissemination Power

The ideological risks brought about by the rapid development of LLMs are, in essence, the double-edged sword effect resulting from technological development, and the product of continuous breakthroughs and iterations of large models in technological dimensions such as data and algorithms. Therefore, preventing the ideological risks they pose equally requires returning to the technological substrate for regulation. In the technological dimension, centering on the problem domains of political security, value security, and public opinion security of generative large language models, the negative effects of instrumental rationality must be mitigated from the dimension of value rationality.

First, regarding the technological regulation of political security. Aiming at the illusion of "technological neutrality" and the "post-truth" problem of deepfakes, on the one hand, we shall promote the pre-embedding of political orientation from the generation result to the generation process. Before model training, a sample screening process that reflects political judgment and adheres to fairness and justice should be constructed. Systematic investigation and elimination of potential political biases and ideological hidden dangers in training samples should be carried out, and positive value guidance norms should be clarified, prompting the model to form a distinct, realistic stance during the learning phase. On the other hand, the model must be pushed to shift from generation fluency to ensuring the reliability of generated content. This involves connecting to real-time, authoritative external knowledge bases for output verification and embedding technologies such as factual checking and digital watermarking during the generation process. Thereby, a defense line of facts is constructed within the probability-driven generation mechanism, using the certainty of technology to counter the deceptiveness of deepfakes, laying the foundation for subsequent legal regulation and public identification.

Second, regarding the technological regulation of value security. Focusing on common issues such as data bias and technological black boxes, on the one hand, we must eradicate the output biases caused by data bias, deeply

cultivate the construction of Chinese digital corpora, and break through the ideological biases in large models. This entails expanding the spatial proportion weight of Chinese corpora and utilizing the "space-time folding" mechanism to expand the temporal proportion weight of Chinese corpus content from the past decade. Using the cultural symbols of Chinese civilization in the two-dimensional field of space and time to implement value-oriented interventions in artificial intelligence will help embed the mainstream socialist ideology into the system core of the model, achieving high-quality and high-volume content production. On the other hand, the structural problems of the algorithmic black box must be solved by advancing the construction of interpretability tools. First, construct interpretability oriented towards "hidden layers," building a visualization model capable of revealing the causal relationship between data inputs and procedural outputs. Second, construct interpretability oriented towards the individual dimension. Build a realistic communication bridge between technology developers and individual users. In the conceptual explanation dimension, use comprehensible, popular language to explain the basic theories relied upon for the generation of procedural data results. In the data analysis dimension, explain the relevant data and its proportion weights upon which the output content is based, using text or image narratives. In the correlation explanation dimension, clarify the presence or absence of causal relationships and correlations between contents, and between data points, for the output content.

Third, regarding the technological regulation of public opinion security. Focus on the social problems triggered by "immersive narratives" and "emotional arousal." On the one hand, construct a consistency review mechanism to regulate the social biases of audiences. Based on the establishment of independent, controllable, high-quality, and standardized databases, set up a factual consistency review process for the narrative content generated by models. Verify the consistency between the generated text information and social reality, identify and filter out content contrary to objective facts, and resist the dissolution of rational thinking caused by excessive immersion. On the other hand, strengthen the construction of semantic understanding capabilities and regulate the cognitive guidance of audiences. Rely on semantic analysis tools to systematically analyze the emotional trends, narrative directions, and value orientations of generated content. Accurately identify emotionally arousing content such as "emotionally engaging narratives with drifting values" or "emotional incitement lacking positive guidance." (Zhao & Chen, 2025) Systematically filter out extreme, exclusive, and overly emotional "tribal" language therein, in order to suppress the model's deliberate guidance and amplification of public opinion from the source.

4 Conclusion

When an emerging technology appears, it is often accompanied by a dichotomy of coexistence between acceptance and rejection. Observing the current generative large language model technology, it is both enthusiastically embraced as a revolutionary tool that disrupts efficiency and widely feared due to the hidden dangers it carries in disseminating ideological risks. A systematic analysis of the ideological attributes and potential ideological risk issues of generative large language models actually directs the problem domain towards how to reconcile this opposition between "acceptance" and "rejection," promoting technological development to advance in tandem with the public interests of society. Hubert Dreyfus pointed out: "The progress of technology presents a danger. This danger lies not in specific technological advances or technological tools, but in our understanding of ourselves, in the enlightenment we derive from a technological way of life." (Dreyfus & Kelly, 2014: 207) As the intellectual fruit of human labor practice, generative large language models originate from human beings and must ultimately serve human beings.

Their development should always take technological innovation serving the enhancement of human subjectivity and the manifestation of social publicness as its fundamental orientation (Yang, 2025). Only in this way can technology be truly validated as the realization of the "essential powers of man," thereby avoiding becoming a new type of ideological carrier that alienates social relations.

References

[1] Chinadaily.com.cn. (2024, July 22). *Resolution of the Central Committee of the Communist Party of China on further deepening reform comprehensively to advance Chinese modernization*. https://www.chinadaily.com.cn/a/202407/22/WS669db327a31095c51c50f2f8.html.

[2] Cyberspace Administration of China. (2023, October 18). *Global AI governance initiative*. https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm

[3] Dreyfus, H., & Kelly, S. D. (2014). *All Things Shining* (J. Q. Tang, Trans., p. 207). Shandong Literature and Art Publishing House.

[4] Jiang, J. M. (2024). Challenges and coping strategies for the dissemination of mainstream ideology under the influence of recommendation algorithms. *Studies on Marxism,* (3), 143-154+156.

[5] Li, R. Q., & Li, G. L. (2025). The ideological risks of generative artificial intelligence: Appearance, patterns, and governance. *Contemporary World and Socialism,* (2), 149-156.

[6] Liu, Y. J. (2010). On the construction of ideology by science and technology. *Social Science Front,* (5), 240.

[7] Marx, K., & Engels, F. (2009a). *Collected Works of Marx and Engels (Vol. 2)*. People's Publishing House.

[8] Marx, K., & Engels, F. (2009b). *Collected Works of Marx and Engels (Vol. 3)*. People's Publishing House.

[9] Marx, K., & Engels, F. (2009c). *Collected Works of Marx and Engels (Vol. 5)*. People's Publishing House.

[10] Marx, K., & Engels, F. (2009d). *Collected Works of Marx and Engels (Vol. 8)*. People's Publishing House.

[11] Miao, Q., Wang, X., Yang, J., Zhao, Y., Wang, Y., Chen, Y., Tian, Y., Yu, Y., Lin, Y., Yan, R., Ma, J., Na, X., & Wang, F. (2024). From foundation intelligence to general intelligence: The state-of-the-art and perspectives of GenAI and AGI based on foundation models. *Acta Automatica Sinica, 50*(4), 674–687.

[12] Shang, H. P., & Liu, H. M. (2025). The Western ideal of "value neutrality of technology" and the "apologetic" paradox in national governance usage: An exploration of the practical use of ChatGPT technology. *Academic Monthly, 57*(9), 76-89.

[13] Wang, X. N. (2025). Integrating Value Consensus through Technology Symbiosis: Grasping the dividends of smart technology to promote the upgrade of values education. *e-Education Research, 46*(1), 27.

[14] Wu, X. Q. (2024). From ChatGPT to Sora: An examination of the "truth" in ideological discourse. *Ideological Education Research,* (8), 75-83.

[15] Xi, J. P. (2018, November 1). Xi stresses strengthening leadership, planning, task definition and foundation laying to promote sound development of new-generation artificial intelligence in China. *People's Daily*, p. 1.

[16] Xiang, J. Y. (2025). The ideological risks of text-to-video artificial intelligence and its prevention and control paths. *Jianghan Tribune,* (8), 67-74.

[17] Yang, X. T. (2025). The ideological metaphor of generative artificial intelligence and the response to its reified risks. *Studies on Marxism,* (9), 133-144.

[18] Zhang, T. (2025). The construction of historical authenticity by large language models, its risks, and countermeasures. *Tianjin Social Sciences,* (6), 28-37.

[19] Zhao, X. Q., & Chen, X. M. (2025). Ideological risks of text-to-video models and strategies for their prevention and resolution. *Journal of South China Normal University (Social Science Edition)*, (5), 194-206+208.

[20] Zhou, M. P., & Jin, X. (2024). Challenges and alleviation of college students' mainstream ideological identity from the perspective of intelligent algorithm application. *Hunan Social Sciences,* (3), 146-153.

[21] Zhu, W. J. (2026). The trust crisis and reconstruction of generative artificial intelligence. *Journal of Tianjin Administrative Institute, 28*(1): 41-53.