Proceedings of the 5th International Conference on Literature, Language, and Culture Development
DOI: 10.54254/2753-7064/2026.35420

Government, Enterprises, and Public Opinion: Influencing Factors Driving the Evolution of AI Value Alignment Agenda

Zhiwen He

School of Global Journalism and Communication, Southwest University of Political Science and Law, Chongqing, China
Hijikatatoushirou0505@outlook.com

Abstract. Since the release of ChatGPT in November 2022, generative artificial intelligence (AIGC) has rapidly penetrated all fields of civil applications. This process has broken the traditional paradigm that AI research, development and feedback were confined to laboratories, turning the vast public opinion arena into a critical testing ground for AI value alignment. This study focuses on the critical policy window from March 2023 (concurrent release of Ernie Bot and GPT-4) to August 15 (official implementation), and examines how public opinion, through the mechanism of "reverse agenda setting", cooperates with government regulation to promote AI value alignment and technological adaptation. Using co-occurrence network analysis and quadratic assignment procedure (QAP) regression model, this study empirically tests the keyword networks of Sina Weibo users, official accounts of AI enterprises and policy texts of the Cyberspace Administration of China. The findings show that: Public forcing mechanism: In the early stage of policy making, public opinion significantly predicts the government agenda ($\beta=0.216$, $p<0.001$). Limitations of government regulation: The government transforms fragmented public opinion into structured policies by screening issues such as "security". Tripartite governance paradigm: After government intervention, the explanatory power of the model is significantly improved ($R^2=0.488$), forming a closed loop of "public initiation—government regulation—enterprise framing". This study theoretically expands network agenda setting (NAS) theory by extending the research subject from the traditional "media-public" duality to the "public-government-enterprise" tripartite interactive structure.

Keywords: Public Opinion, AI Value Alignment, Network Agenda Setting, Reverse Agenda Setting, AI Risk Management

1. Introduction

1.1. Research background

Agenda-setting theory has developed for half a century, evolving from the first level exploring the salience of media issues to the second level examining attribute agendas, and in recent years has developed into network agenda setting (NAS) theory centered on "associative networks". This evolutionary track has closely followed the iteration of media technology. At present, as generative

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.35420

artificial intelligence permeates social life at an unprecedented speed, AI technology itself has become a key variable reshaping the communication ecology. This not only requires algorithms to align with human values, but also calls for communication theory to respond in the new context: when AI is both a communication carrier and a communication subject, how will the traditional logic of agenda setting change?

Meanwhile, the phenomenon of reverse agenda setting has heralded the rise of the public as a communication subject since the 1990s [1]. In March 2023, with the successive release of Ernie Bot and GPT-4, the Chinese Internet witnessed a heated discussion on AI ethics, risks, and future. In this process, the public is no longer just a passive recipient of policy regulation, but has constructed a "cognitive schema" of AI risks through social media. Notably, this bottom-up agenda flow did not spread disorderly; it was particularly active during the incubation period (March–April 2023) of the Interim Measures for the Management of Generative Artificial Intelligence Services (hereinafter referred to as the Interim Measures), showing obvious characteristics of "public opinion forcing" [2]. However, different from the traditional "media-public" dual game, reverse agenda setting in the field of AI governance must go through the government's "gatekeeping" and "translation"—that is, refining fragmented public emotions into structured policy red lines—to affect the technical end truly.

1.2. Research purposes

Existing agenda-setting research mostly focuses on linear interactions between two subjects, rarely paying attention to whether the intervention of a third subject reshapes the path and effectiveness of agenda flow in complex technical governance scenarios. Based on this, this study takes March to August 2023 as the observation window and selects Sina Weibo as the observation field of public opinion, aiming to clarify the following core questions:

First, test whether the keyword network structure of public opinion has predictive power for government policy texts before and after the issuance of the Interim Measures, that is, verify the existence and boundaries of "public reverse agenda setting".

Second, deeply analyze the role of the government in the "public-enterprise" dual structure. This study attempts to break away from the binary opposition of "unilateral government dominance" or "passive government response", and explore how the government functions as a "regulatory hub" between absorbing public demands and regulating enterprise development.

Finally, this study is committed to constructing a tripartite interactive model including "public-government-enterprise", revealing the complex game and coupling mechanism among public opinion emotion flow, policy regulation flow, and technological business flow in the process of AI value alignment.

2. Literature review and research significance

2.1. Technical ethics perspective: sample bias and local deficiency in AI value alignment

The core challenge of AI value alignment lies in translating abstract "human values" into specific algorithmic reward models (RLHF) . Existing studies show that global AI training data generally suffers from "Western-centrism" bias, with a severe shortage of social ethics samples in non-Western contexts, making it difficult for AI output values to conform to local social expectations [3]. In addition, due to the lack of cross-cultural longitudinal studies, the academic community still knows little about the dynamic evolution mechanism of public attitudes toward AI [4]. This means that

DOI: 10.54254/2753-7064/2026.35420

relying solely on laboratory data cannot achieve true universal alignment, and there is an urgent need to introduce large-scale, localized public opinion data as a supplement, which is the gap this study attempts to fill.

2.2. Governance transformation perspective: dilemma of government agenda setting in the internet era

Under the traditional governance model, the government can dominate the social agenda by regulating mainstream media [5]. However, in the self-media era, the decentralized communication structure has weakened the effectiveness of traditional control, leading the government to often face dilemmas of "insufficient response" or "regulatory failure" in public opinion guidance [6]. Although some scholars have proposed digital governance models such as "three-dimensional collaboration", most of these models are not specially designed for the high risk and uncertainty of AI technology [7]. Especially under China's unique top-down pressure mechanism, the government often falls into the cycle of "passive response" rather than "active adaptation". Therefore, exploring how the government can effectively coordinate public emotions and technological development in AI governance is a realistic proposition that urgently needs to be addressed by current agenda-setting theory.

2.3. Communication mechanism perspective: absence of "government regulation" in reverse agenda setting

Reverse agenda-setting theory confirms that the public has the ability to reconstruct the media agenda through self-media. However, existing reverse agenda research mostly focuses on the "public-media" dual interaction, ignoring the filtering and translation functions of the "government" as a key node in complex public affairs [8]. Studies show that public agendas are often emotional and fragmented, and it is difficult to directly transform into governance effectiveness without the "issue separation" and screening of authoritative subjects . In the AI governance field, although the public provides "raw materials" for risk early warning, how the government gatekeeps, selects and structurally reconstructs these scattered keyword networks and then transmits them to the enterprise end remains an under-revealed black box.

3. Research methods

3.1. Research design and data sources

This study selects March 2023 (release of Ernie Bot and GPT-4) to September 2023 (post-implementation period of the Interim Measures) as the research window, focusing on two stages: policy incubation (March–April) and implementation (March–September) . The data cover three major subjects:

Public opinion: 32,000 ordinary user posts containing keywords "AI/artificial intelligence" on Sina Weibo;

Enterprise discourse: More than 3,100 posts published by official accounts of mainstream large models such as Ernie Bot and Tongyi Qianwen;

Policy texts: More than 4,100 policies and reports issued by the official website of the Cyberspace Administration of China, People's Daily, and CCTV News.

DOI: 10.54254/2753-7064/2026.35420

3.2. Text processing and co-occurrence matrix construction

To capture the semantic association structure between keywords, this study adopts the sliding window method (window=5 sentences) to construct word pair co-occurrence matrices.

Preprocessing: After denoising, Jieba word segmentation, and case unification, an initial matrix is generated.

Binarization processing: To focus on network structure similarity and weaken frequency differences, the matrix is converted into a 0-1 binary matrix (1 for co-occurrence, 0 otherwise) .

Threshold screening: To solve the failure of the QAP test caused by an over-dense network (density≈1), the quantile threshold method is adopted, retaining only the top 20% of edges with co-occurrence strength in each matrix, so that the network density is stable in a reasonable range of 0.15–0.25.

Staged matrices: "March–April public/government matrices" and "March–September public/government/enterprise matrices" are produced according to time nodes, corresponding to the policy incubation period and the full cycle, respectively.

3.3. Model setting and testing

This study uses the quadratic assignment procedure (QAP) for regression analysis, correcting standard errors through 5,000 permutation tests to overcome the pseudo-regression problem caused by network autocorrelation .

Stage 1: Testing public reverse agenda setting

Taking the April CAC policy matrix as the dependent variable (Y) and the March Weibo public opinion matrix as the independent variable (X), test whether the public agenda significantly predicts the policy agenda.

$$\mathrm{YGov}\,(\mathrm{Apr}) = \beta 0 + \beta 1\ \mathrm{XPublic}\,(\mathrm{Mar}) + \varepsilon \tag{1}$$

Stage 2: Testing the positive allocation effect of the government

Taking the full-cycle public agenda as the dependent variable, incorporating the full-cycle government policy matrix and enterprise matrix as independent variables, construct a multivariate QAP model to test the evolution of the agenda structure after government intervention.

$$\mathrm{YPublic}\,(\mathrm{Full}) = \beta 0 + \beta 1\ \mathrm{XGov}\,(\mathrm{Full}) + \beta 2\ \mathrm{XFirm}\,(\mathrm{Full}) + \varepsilon \tag{2}$$

Auxiliary analysis: Visualize the above five matrices, compare the dynamic evolution of core node centrality and community structure, and corroborate the agenda flow path.

4. Results

4.1. Structural evolution characteristics of keyword co-occurrence networks

Through topological structure analysis of keyword co-occurrence matrices under different time windows, this study finds that the semantic network of AI issues shows significant staged differentiation and recombination characteristics.

First, in the policy incubation period (March–April), the network structure presents the dislocation and connection of "risk perception" and "governance preposition". The public opinion network has a relatively high density (Density≈0.21), with its core nodes highly concentrated in

DOI: 10.54254/2753-7064/2026.35420

negative emotion and risk perception words such as "unemployment", "privacy leakage", "algorithmic discrimination", and "out of control", forming a typical "anxiety-driven" network structure . This indicates that under the impact of the dual release of Ernie Bot and GPT-4, the public's primary concern is the existential threat and ethical challenges brought by AI. In contrast, the government policy network in the same period (Density≈0.18), although also involving "risk" issues, has its core nodes rapidly shifted to technical governance terms such as "compliance", "filing", "security assessment", and "generative". This structural difference suggests that the government did not passively wait for public opinion to ferment, but quickly launched a "governance preposition" mechanism, attempting to transform public emotional risks into operable institutional red lines.

Second, at the full-cycle (March–September) level, the network structure evolves into a stable form of "tripartite co-governance". With the implementation of the Interim Measures, the centrality of the public network has shifted significantly: the centrality of risk-related words has decreased, while pragmatic words such as "application scenarios", "efficiency improvement", and "AI assistant" have risen, showing that the public's mindset has turned from panic to adaptation. The enterprise network has always maintained a technological and business closed loop centered on "large model", "computing power", "ecology", and "Qianfan (Baidu Smart Cloud)", with the lowest overlap with other networks, showing extremely strong agenda independence. Notably, government network nodes are always in the "betweenness" position in the visual map, connecting the public's "risk side" and the enterprise's "technology side", playing an indispensable role as a bridge and valve.

4.2. QAP regression yest results

To accurately quantify the causal relationship between the above network structures, this study performed quadratic assignment procedure (QAP) regression analysis, and the results are shown in Table 1 .

Table 1. Summary of QAP regression analysis results

Variable Type	Stage 1: Public→ Government (Mar–Apr)	Stage 2: Full-factor Model (Mar–Sep)
Independent Variables	β (Std.)	β (Std.)
Public Agenda (Weibo)	0.216*	0.601*
Government Agenda (CAC)	—	0.208***
Enterprise Agenda (AI Manufacturers)	—	0.195***
R^2	0.047	0.488
Adjusted R^2	0.046	0.487
Permutation Test (p)	< 0.001	< 0.001
Number of Network Nodes (N)	312	315

Note: *p < 0.05, **p < 0.01, ***p < 0.001

The in-depth analysis of the results is as follows:

Stage 1: Trigger effect and limitations of public reverse agenda

DOI: 10.54254/2753-7064/2026.35420

Model 1 data show that the keyword network structure of public opinion in March has a significant positive predictive power for government policy texts in April ($\beta=0.216$, $p<0.001$). This confirms that on the eve of the public consultation of the Interim Measures, the public successfully "problematized" AI risk issues through social media, thus triggering the government's agenda-setting [9]. However, the R^2 value of the model is only 0.047. Although this value is statistically significant, it is extremely weak in explanatory power. It reveals a key mechanism: the public agenda acts more as an "igniter" or "alarm", providing raw materials and emotional signals for the issue, but far from a complete template for the policy blueprint. The government did not fully accept public demands, leaving huge room for subsequent screening and reconstruction.

Stage 2: Explanatory power leap of tripartite interaction

Model 2 introduces government and enterprise variables, and the results undergo qualitative changes. The standardized coefficient of the public agenda jumps sharply to 0.601, and both the government ($\beta=0.208$) and enterprise ($\beta=0.195$) agendas show significant positive effects. Most importantly, the R^2 of the model soars from 0.047 to 0.488, meaning that the model's explanatory power for agenda network variation increases more than tenfold [10]. This data strongly supports the "tripartite interaction" hypothesis of this study:

Strong dominance of public opinion (0.601): After the government establishes rules, public emotions become the "weather vane" of the market. To obtain commercial legitimacy, enterprises must frequently respond to public concerns (such as data security and ethical bottom lines) at the discourse level, resulting in an overwhelming impact of the public agenda on the enterprise agenda.

Strategic regulation of the government (0.208): The government coefficient is significant but weaker than that of the public, confirming its positioning as a "regulatory hub" rather than a "daily driver". The government's role is not to guide every topic in detail, but to draw insurmountable red lines through core frameworks such as "security" and "controllability".

Selective framing of enterprises (0.195): As an independent agenda source, enterprises attempt to guide the public's fear of risks into expectations for technological dividends under the premise of compliance by emphasizing frameworks such as "technological empowerment" and "industrial ecology", reflecting the implicit discourse power of enterprises.

4.3. Establishment of the "tripartite interaction" narrative model

Integrating network structure evolution and regression data, this study constructs and verifies the unique "public initiation—government regulation—enterprise framing" tripartite interaction narrative model in the AI governance field. This model breaks through the binary opposition framework of "media-public" in traditional agenda setting. The study finds that a single public or government subject model has a serious explanatory power deficit (very low R^2), and only when the three form a closed loop can the complex picture of AI value alignment be fully explained. In this model, the public is the "trigger point" of the issue, responsible for inputting social risk preferences; the government is the "filter" of values, responsible for sublimating fragmented public opinion into structured public policies; enterprises are the "interpreters" of technology, responsible for finding a balance between policy implementation and market catering. The three are interdependent and mutually restrictive, jointly constituting the dynamic evolution path of value alignment in the AI era.

DOI: 10.54254/2753-7064/2026.35420

5. Discussion

5.1. Liminality of reverse agenda: structural leap from emotional catharsis to policy issues

Taking March to April 2023 as the window, this study empirically proves that public opinion has significant predictive power for the government policy agenda ($\beta=0.216$, $p<0.001$), confirming the existence of reverse agenda setting in the network era [11]. However, in-depth analysis of this micro-mechanism reveals that the public agenda has significant "liminality" when touching the core of policies .

Data show that the public agenda acts more as a "risk early warning device". During the policy incubation period, the keyword network in the Weibo arena presents highly "emotional" and "generalized" characteristics, with core nodes such as "unemployment", "privacy leakage", and "Terminator" mainly reflecting the public's instinctive fear and ethical anxiety about unknown technologies. Although this agenda flow has strong mobilization ability enough to "open the door" to policies, it lacks the technical feasibility and institutional flexibility required to be transformed into legal provisions.

Therefore, the coefficient of 0.216, together with its weak R^2 value (0.047), reveals a fact ignored by traditional reverse agenda theory: in the AI governance field with high technical barriers, the public has the ability to "raise questions" but often lacks the power to "define questions". The public agenda must undergo a qualitative change from "emotion flow" to "policy language" to take effect. This finding revises the romantic imagination of "Internet empowerment" in early studies, pointing out the structural shortcomings of the public in complex technical governance—that is, the "initiation right" of the agenda is not equal to the "decision-making power".

5.2. Reconstruction of government role: "regulatory hub" and strategic translation

Combined with the significant coefficient of the government agenda in Stage 2 data ($\beta=0.208$), this study proposes a new government role positioning of "regulatory hub", distinguishing it from the traditional "unilateral dominance" or "passive response" model.

Empirical results show that the government is not a simple mirror of public demands, but plays an extremely critical role of "strategic translator" [12]. Specifically, the government carries out "issue separation" on the public agenda through a sophisticated screening mechanism:

Absorption and elevation: For issues in line with national governance goals, such as "data security" and "algorithmic discrimination", the government abstracts them from specific social complaints into national policy languages such as "controllable development" and "ethical norms", and solidifies them into mandatory institutional red lines through the Interim Measures.

Suppression and guidance: For overly radical, irrational demands in public emotions or those beyond the current technical governance capacity (such as "suspending AI R&D"), the government conducts cold treatment or reframing through the agenda setting of mainstream media to prevent unnecessary social panic.

Notably, the government coefficient (0.208) is significantly lower than the public coefficient (0.601) in the full model, which strongly proves that the government is not the "driver" of daily AI discourse, but the "traffic commander". Its role is not to guide every topic in detail, but to channel the "turbulent flow" of public emotions into the "main channel" of technological innovation by drawing the bottom lines of "security" and "compliance". This is a more mature "dredging governance", that is, while acknowledging the public's right to know, maintaining the country's strategic determination in technological evolution.

DOI: 10.54254/2753-7064/2026.35420

5.3. Implicit hegemony of enterprises: "selective framing" and embedded business logic

Another core breakthrough of this study is to reveal the unique operational logic of enterprises as a third subject in the agenda network (β=0.195). Data visualization and regression result jointly point to a fact: enterprises implement a systematic "selective framing" strategy when responding to external pressures.

Faced with public risk anxiety and government compliance requirements, enterprises do not passively conduct defensive public relations but take the initiative to reconstruct the semantic field of dialogue [13]. By positioning themselves as "technological enablers" and "efficiency improvers", enterprises frequently output frameworks such as "industrial ecology", "cost reduction and efficiency improvement" and "new quality productive forces", cleverly replacing the public's fear of "existential threat" with expectations for "technological dividends".

This "selective framing" is essentially an implicit hegemony of "tech-business collusion". Using its algorithmic architecture power, enterprises preset the technical path of AI development, forcing the public and the government to only discuss "how to use AI" without touching the fundamental question of "whether to want AI". This strategy not only dissolves part of regulatory pressure but also subtly packages "maximizing commercial interests" as a synonym for "social progress". In the practice of AI value alignment, if only public supervision and government regulation are focused, while ignoring the invisible manipulation of the agenda by enterprises as "architectural power" holders, the so-called "human alignment" is likely to be alienated into "capital alignment".

5.4. Practical implications and policy recommendations

Based on the above findings, to achieve truly "people-oriented" AI governance, this study puts forward the following suggestions: For the government: An "agile perception-dynamic translation" mechanism should be established. It is necessary not only to monitor public opinion heat but also to have the "translation ability" to transform vague public demands into precise policy tools. Meanwhile, vigilance should be maintained against enterprises using technological hegemony to evade regulation, and substantive review of algorithm filing and data annotation links should be strengthened. For enterprises: Thinking beyond "compliance defense" should be adopted, and the initiative to assume the responsibility of "algorithmic transparency" should be taken. Instead of only diverting attention through "selective framing", a normalized public communication channel should be established, and social ethics should be embedded in the whole life cycle of technological R&D, not just as a public relations strategy. For the public: It is necessary to improve one's own "algorithmic literacy". Attention should be paid not only to the short-term risks brought by AI but also to the long-term commercial incentives under the packaging of enterprise technical discourse, and participation in the formulation of AI rules through rational public discussion.

6. Conclusion

6.1. Research summary: from binary opposition to "tripartite spiral"

Taking the policy window of generative AI governance in China in 2023 as a slice, this study systematically investigates the interaction mechanism among the public, the government, and enterprises in the process of AI value alignment through the QAP regression model and semantic network analysis. The study confirms the effectiveness of the tripartite spiral model of "public initiation—government regulation—enterprise framing".

DOI: 10.54254/2753-7064/2026.35420

Data show that although public opinion can trigger the policy agenda through reverse agenda setting (β=0.216), its fragmented characteristics determine that it must take effect through government "translation". In this process, the government plays a key role as a "regulatory hub" (β=0.208), transforming emotional risk perception into structured institutional red lines. Under compliance pressure, enterprises guide public attention to technological dividends through "selective framing" (β=0.195), thus seeking a balance between business logic and social responsibility. The closed loop formed by the three has an explanatory power (R^2=0.488) far exceeding that of the traditional dual model, revealing the complexity of power flow in the AI governance field.

6.2. Theoretical contributions

This study revises and expands the existing theoretical system in the following three aspects: First, expansion of the subject of agenda-setting theory (NAS). Breaking through the "media-public" dual framework long focused on by McCombs et al., this study establishes "enterprises" as the third pole agenda subject following the government and the media. In the algorithm society, enterprises master the "architectural power" at the code level, and their "selective framing" through technical black boxes essentially constitutes an implicit and imperceptible agenda setting, which is an important supplement to classic communication theory. Second, revision of the "responsive governance" paradigm. Existing studies mostly regard the government as a passive "pressure receiver". The "regulatory hub" model proposed in this study shows that the government is not simply "passively responding" in AI governance, but a "strategic active adaptation". Through screening the public agenda (absorbing "security" and suppressing "panic") and guiding enterprise behavior, the government has realized the creative transformation from "emotion flow" to "governance flow", enriching the theoretical connotation of public management in the digital era. Third, reinterpretation of the path of AI value alignment. Breaking away from the purely technical perspective (such as RLHF), this study points out that AI value alignment is essentially a triple game of "national governance logic", "commercial capital logic", and "social public logic" in practice. Public demands must go through the double filtering of "government filtering" and "enterprise translation", which provides a new empirical clue for understanding the political economy dimension of AI ethics.

6.3. Research limitations and future prospects

This study still has the following limitations, which are also the directions for in-depth research in the future: First, sample and time constraints. This study focuses on the period of drastic policy changes, and the applicability of the conclusions under normal governance needs to be tested. In the future, the observation cycle can be extended to compare cultural differences in AI governance agendas between different countries (such as China, the United States, and Europe). Second, the black box of "word-action" transformation. This study is mainly based on text semantic networks. Although it can reveal the agenda structure, it cannot fully restore the specific administrative process from "Internet hot words" to "official documents". In the future, computational ethnography or interview methods can be used to deeply describe the cognitive schema of policymakers. Finally, the challenge of technological iteration. With the maturity of Agent technology, AI itself may become a new agenda-setting subject. Future research can introduce large language model Agents to simulate the evolution of public opinion and explore new laws of agenda setting in the era of human-machine symbiosis.

DOI: 10.54254/2753-7064/2026.35420

References

[1] Jaung, W. (2026). Does AI value the environment? Evaluation of AI value alignment. *Technological Forecasting & Social Change, 225*, 124550.
[2] Novotny, M., Weber, W., Kern, C., & Kreuter, F. (2025). Measuring public opinion towards artificial intelligence: Development and validation of a general AI attitude short scale. *AI & Society*. Advance online publication. 1–33.
[3] Qi, W., Pan, J., Lyu, H., & Luo, J. (2024). Excitements and concerns in the post-ChatGPT era: Deciphering public perception of AI through social media analysis. *Telematics and Informatics, 92*, 102158.
[4] Moriniello, F., Martí Testón, A., Muñoz, A., Silva Jasaui, D., Gracia, L., & Solanes, J. E. (2024). Exploring the relationship between the coverage of AI in WIRED magazine and public opinion using sentiment analysis. *Applied Sciences, 14*(5).
[5] Zhang, Y. (2026). Challenges and optimization paths of media discourse in the context of we-media. *Journal of News Research, 17*(4), 14–19.
[6] Wen, H. (2025). The impact of we-media on the news communication pattern of traditional media. *China Newspaper Industry*, (18), 64–65.
[7] Liu, M., & Liu, L. (2025). Research on government public opinion guidance mechanism and credibility construction in the new media environment. *International Public Relations*, (20), 152–154.
[8] Xu, P. (2024). Analysis of reverse agenda setting from the perspective of actor-network theory: Taking the "landlord inspecting damage with a lamp" incident in Jiangxi as an example. *Journalism & Communication*, (18), 25–27.
[9] Triantantyllopoulos, L., Paxinou, E., Tzanoulinou, D., Verykios, V. S., & Kalles, D. (2026). The value alignment problem in advisory AI: A systematic literature review. *AI and Ethics, 6*(1), 147.
[10] Huang, L. T. L., Papyshev, G., & Wong, J. K. (2024). Democratizing value alignment: From authoritarian to democratic AI ethics. *AI and Ethics, 5*(1), 1–8.
[11] Watson, E., Viana, T., Zhang, S., Sturgeon, B., & Petersson, L. (2024). Towards an end-to-end personal fine-tuning framework for AI value alignment. *Electronics, 13*(20), 4044.
[12] Peterson, M., & Gärdenfors, P. (2023). How to measure value alignment in AI. *AI and Ethics, 4*(4), 1493–1506.
[13] Schuster, N., & Kilov, D. (2025). Moral disagreement and the limits of AI value alignment: A dual challenge of epistemic justification and political legitimacy. *AI & Society, 40*(8), 1–15.