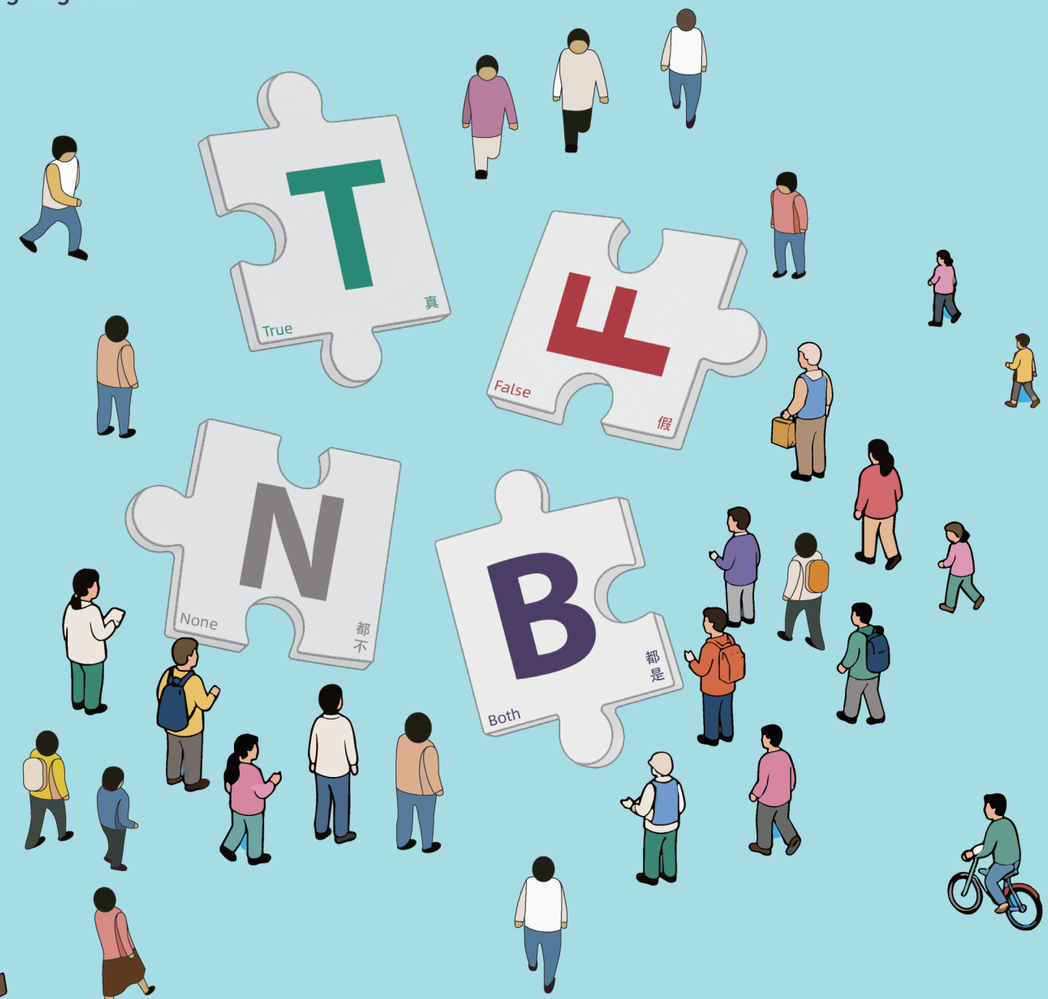
FUDAN
REPORT SERIES

WAIC 2026 Forum
AI Global Governance and Sustainable Development: Challenges and Countermeasures

Risk Transformation and Ethical Responses in the Age of ASI

Yang Qingfeng et al.



Center for Global AI Innovative Governance
Institute of Technology Ethics for Human Future

CGAIG 全球人工智能创新治理中心
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

Fudan Think Tank Building,
No. 220 Handan Road, Shanghai, China, 200433
Tel: +86-21-65641067
Email: cgaig@fudan.edu.cn
Website: https://cgaig.fudan.edu.cn

FUDAN REPORT SERIES

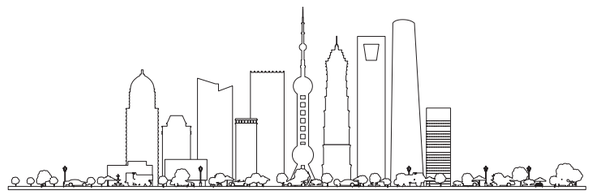
Risk Transformation and Ethical Responses in the Age of ASI

Yang Qingfeng et al.

Center for Global AI Innovative Governance (CGAIG)

Institute of Technology Ethics for Human Future, Fudan University

June 2026



Contents

Key Points

01 At present, AI ethics exhibits the following characteristics: its academic orientation toward the pursuit of the laws of freedom is gradually weakening, while tendencies toward legalization, institutionalization, and engineering are becoming increasingly prominent.

02 AI ethics can be categorized into technology-oriented and humanistic-oriented ones. Its development needs to shift from a "product" mindset to a "practice" mindset.

03 Technology-oriented AI ethics includes ethical issues related to technological scenarios, technical challenges, technological capabilities, and the entire life cycle of AI systems.

04 Humanistic-oriented AI ethics includes ethical issues related to hermeneutics, humanistic critique, and human-machine relations.

05 ASI risk manifests as a spherical risk of an existential nature.

06 In the age of ASI, the ethics of human-machine relations should move beyond the binary opposition between humans and machines. Establishing ethical contracts between humans and robots is a feasible approach. Dynamic contractual mechanisms may be employed to monitor ethical frictions, while conflicts suspended under four-valued logic can ultimately be resolved through human deliberation.

07 In the future symbiotic society of humans and machines, intelligent contracts grounded in the theory of affirmative contracts may constitute a viable path, and "Ethicist AI" represents a practical attempt at implementing such intelligent contracts.

Abstract

Ethics is commonly understood as a discipline primarily concerned with uncovering the laws of freedom and constructing normative principles. Yet fulfilling this mission also reveals ethics' inherent fragility. Compared with law and institutional systems, ethical principles lack rigid binding force. Consequently, AI ethics, which likewise aims to establish normative principles, exhibits this same inherent fragility. With the accelerated iteration of AI technologies and the approach of the age of Artificial Super Intelligence (ASI), AI ethics has gradually undergone three major transformations: institutionalization, legalization, and engineering, all of which are intended to overcome this inherent fragility.

As ASI increasingly shifts from technological imagination to a real-world problem, the legalization, institutionalization, and engineering of AI ethics alone remain unequal to address its challenges. ASI differs markedly from Artificial General Intelligence (AGI): it is emerging as a new subject restructuring productive forces and relations of production, and super-agency has become the key concept shaping the human-AI relationship. Taking AI risk as an example, relevant theories are grounded in the hierarchical classification of hazards posed by AI agents, notably the EU AI Act, which classifies systems into four categories: minimal, limited, high, and unacceptable risk. Nevertheless, technical standards for high-risk and unacceptable-risk AI systems have failed to materialize as scheduled.① ASI, however, transforms existence, creating a new existential structure. Drawing on Peter Sloterdijk's Spherology, this report demonstrates that AI risk has evolved into an ontological structural

① A growing disconnect has emerged between the AI Act's risk-based regulatory approach and industrial realities. To prevent the Act's de facto unenforceability due to unready harmonized standards, high-risk obligations' enforcement should be dynamically tied to these standards' availability. Harmonized standards serve as the official bridge translating abstract AI ethical principles into auditable, implementable technical requirements. On 20 November 2025, the European Commission introduced the Digital Omnibus, creating a dynamic mechanism that links the entry into force of the high-risk system rules to the availability of technical standards, with a set of buffer and supporting measures.

problem, manifesting as complex spherical risks. In response to these challenges, the construction of AI ethics must achieve a paradigm shift from a "product" mindset to a "practice" mindset. The new AI ethics framework comprises two key dimensions: technology-oriented ethics, focusing on ethical issues arising from critical pain points in application scenarios, competence envelope, and full life cycle; and humanistic-oriented ethics, addressing ethical concerns related to humanistic interpretation, critique of technology, and human-machine relationship reconfiguration. Against ASI-induced power imbalances, this report proposes an "Intelligent Contract Ethics" framework that reconstructs human-machine relations through contractual ethics, thereby securing a digital grounding for human subjectivity. In engineering practice, the framework employs dynamic contracts and mechanisms to monitor ethical frictions in real time, temporarily suspend conflicts when they arise, and submit them to human deliberation via a four-valued logic system, thereby ensuring humanity's ultimate decision-making authority. Currently, contract ethics is shifting from unidirectional alignment to bidirectional "co-contracting". For future human-machine symbiosis society, the "Ethicist AI", with intelligent contracts at its core, should be established as a practical paradigm for contract construction.

The more our morals are refined, the more depraved men become

--Johann Woifgang von Goethe, 1776

Introduction

Today, a wide range of initiatives have emerged to draw attention to the AI safety risks, including government proposals,[①] open letters from research institutions,[②] scientists' statements,[③] global appeals,[④] and numerous research reports. Among these, two stand out. The first is The Global Risks Report 2026, which captures delineates AI's transformation from a frontier technology into a systemic force. It fidentifies AI-related risks as among the global risk categories with the sharpest increase over time and the strongest systemic implications. This analysis highlights the centrality of technical practices and governance.[⑤] The second is the International AI Safety Report 2026 by Yoshua Bengio and his colleagues, which classifies emerging ASI risks into three categories: malicious use, malfunction, and systemic risks. This report focuses on technological dimensions.[⑥] Taken together, these reports operate within a framework of "technological risk governance". We identify three underlying assumptions in current AI risk understandings. The first is the premise of "AI as Continuity", which understands AI as a continuous development process moving from ANI, AGI, and ASI, and on which the entire framework of AI ethics is likewise constructed[⑦]. The second is the premise of "AI as Level", which classifies AI systems into distinct hierarchical categories, as exemplified by Google DeepMind's taxonomy. The third is the premise of "AI as Instrument", which views AI primarily as a tool for empowerment and efficiency enhancement.

AI Continuity Theory conceives these three phases as stages within a single continuum. According to this view, AI development is characterized by four dimensions: the gradual evolution of intelligence forms, the steady enhancement of intelligence levels, the progressive realization of general intelligence, and the emergence of ASI as the ultimate stage. Yet it fails to adequately grasp the discontinuity introduced by ASI.

AI Level Theory conceives AI as an entity that can be classified according to distinct levels of capability and autonomy, as exemplified by Google DeepMind's framework. This taxonomy reflects a broader shift from non-intelligent technical artefacts to increasingly autonomous intelligent systems. Strictly speaking, this framework distinguishes between levels of AGI performance and generality on the one hand, and levels of autonomy in human-AI interaction on the other. In terms of capability, the framework identifies six levels of AGI development: Level 0 (No AI), Level 1 (Emerging), Level 2 (Competent), Level 3 (Expert), Level 4 (Virtuoso), and Level 5 (Superhuman). In parallel, it outlines corresponding levels of autonomy in human–AI interaction, ranging from AI as a Tool and AI as a Consultant to AI as a Collaborator, AI as an Expert, and ultimately AI as an Agent.

The AI Instrumental Theory remains widespread in contemporary society. However, this view is often marked by short-termism and carries inherent risks. The unique ontological risk of advanced AI is often overlooked precisely because it differs fundamentally from that of ordinary tools.

The risks at stake here are ontological risks associated with ASI, including transformations in modes of existence, forms of life, and trajectories of existential evolution. Systemic risks manifest only when AI functions as a systemic force. Thet arise from a fundamental transformation in the structure of being: humans, increasingly thrown into the intelligent systems, come to a specific forms of symbiosis with them.

As AI comes to exhibit "super-tool" characteristics, it increasingly appears as an ontological event, giving rise to risks that are ontological rather than merely instrumental. Zhao Tingyang (2022) argues that an event qualifies as an ontological event insofar as it harbors within itself the starting point of a new problem, constitutes a new origin for human life and thought,

① Global AI Governance Initiative.https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html, The Global AI Governance Initiative issued by the Ministry of Foreign Affairs of the People's Republic of China, 20 October 2023; the Global AI Governance Action Plan released by the Ministry of Foreign Affairs of the People's Republic of China, 26 July 2025.

② Global AI Governance Initiative.https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html, The Global AI Governance Initiative issued by the Ministry of Foreign Affairs of the People's Republic of China, 20 October 2023; the Global AI Governance Action Plan released by the Ministry of Foreign Affairs of the People's Republic of China, 26 July 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. The open letter was launched on March 22, 2023, and as of February 5, 2026, it had gathered over 30,000 signatures.

③ https://ASI-statement.org/zh. The statement was launched on October 22, 2025, and as of February 5, 2026, it had gathered over 130,000 signatures.

④ Global Call for AI Red Lines. https://red-lines.ai/

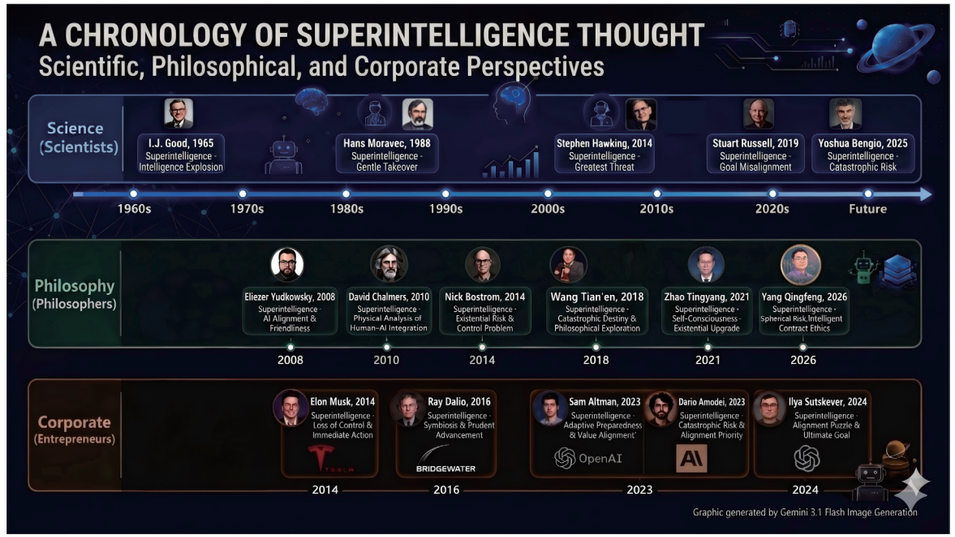
⑤ Global Risks Report 2026. https://www.weforum.org/publications/global-risks-report-2026/

⑥ Bengio, Y et al. International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

⑦ Yang Qingfeng, "Reflections on the Continuity Theory of Artificial Intelligence and an Ontological Exploration", *Social Sciences in China*, 2025, No. 11.

and at root establishes a foundational locus for the mode of human existence[①]. Similarly, Wang Tian'en points out that both the autonomous evolution of human-like intelligence and human-machine evolutionary fusion entail extraordinary ontological consequences.[②]

However, the risks associated with ASI and the corresponding responses to them are far more complex and intractable. They call for Intelligent Contract Ethics, a new form of virtue-oriented ethics suited to the age of ASI. Scientists are increasingly concerned that ASI has gradually evolved from the realm of science fiction into a phenomenon within the scientific world itself.[③] ASI has outgrown its original role as a purely descriptive term for capacities exceeding human abilities. It is increasingly theorized as a phenomenological category in which intelligent agents may be understood as possessing first-personal experience, self-preservation, tendencies, and autonomous modes of action.



Image, produced by Chen Danni

① Zhao Tingyang, "Ontological Events: A View of History", Chinese Social Sciences Today, August 31, 2022, Issue No. 2482. The concept was later revised: a historical event becomes an ontological event if its energy level suffices to transform civilizational structures and modes of existence. Whether an epistemological event that invents technology, knowledge, or thought systems, or a political event that invents institutions, beliefs, or values—so long as it reaches the magnitude of altering civilizational structures, it qualifies as an ontological event. See Zhao Tingyang, "Historicity and Ontological Events", Social Sciences in China, 2023, No. 7, p. 18.

② Wang Tian'en, "The Ontological Implications of Artificial Intelligence", Social Science Front, 2022, No. 5.

③ Yang Qingfeng et al., ASI: The Minority Report, released at the Shanghai Forum in April 2025.

Our previous report *ASI: Minority Report* offered extensive discussions on both technological and philosophical responses to ASI, but provided only a preliminary outline of the ethical response. This year's report seeks to complete that task by proposing more concrete ethical measures.[①]

ASI places humanity within a profoundly complex existential condition. In the past, confronted with scarce natural resources, calamities, and immense survival pressures, nations across the world formed alliances, from which a basic order emerged. This embodies the defining characteristic of a negative contract. Such a contract is established out of necessity; once circumstances change and a particular state becomes sufficiently powerful, it may choose to withdraw from the contract. By contrast, this report is grounded in the intelligent contract, a positive contract conception that we advance.Under this conception, neither human-AI relations bor interpersonal relations remain in a state of nature. From a theoretical perspective, we reflect upon the foundations of traditional contract theory while critically transcending the notion of smart contracts associated with blockchain technology, ultimately establishing a fundamental ethical framework for addressing the relationship between ASI and humanity. From a practical perspective, this contract ethics framework provides the necessary theoretical grounding for the implementation of *Global Digital Compact* (GDC).

Intelligent Contract Ethics is not merely an iidealized framework of our own design. As Johann Wolfgang von Goethe once observed, "The more our morals are refined, the more depraved men become". AI ethics is not merely a set of moral principles articulated by humans; it must also be understood as a normative formation increasingly mediated by intelligent agents.

① Professor Zhang Ping of Peking University has pointed out that addressing the potential risks of ASI and resolving the legal liability dilemmas posed by silicon-based entities have become pressing issues for AI governance. In particular, allocating tort liability for infringements arising from multi-agent collaboration and constructing a risk prevention and control system will prove to be significant challenges.

I. The Evolution of AI Ethics: Trajectories, Emergent Features, and Transformations

Scientists have proposed technical responses to the ontological and catastrophic risks posed by ASI. Among these, Yoshua Bengio's proposal of "Scientist AI" is one of the most influential. In last year's *ASI: Minority Report,* we also outlined ethical responses to ASI, including the idea of "Ethicist AI". Yet these proposals remained largely conceptual and had not been developed into sufficiently standardized or operational forms.

To make this proposal practicable, this chapter undertakes two tasks. First, it reviews the development of AI ethics and clarifies its major trajectories. Second, it articulates the logical and representational conditions under which Ethicist AI and intelligent contracts can be formally constructed. The conditions for this task have become increasingly mature. Within techno-ethics, scholars have begun to reflect critically on the theoretical foundations of AI ethics, while the concepts of techno-ethics are being articulated in diverse ways, including "Ethics as a Service (EaaS)" and "AI safety as a global public good".[①] Mark Coeckelbergh, for example, has rethought established categories to technology ethic through the hermeneutic notion of practice. Building on these discussions, this report argues that AI ethics, if it is to respond adequately to ASI-related risks, must move from a product-oriented mindset to a practice-oriented one. Such a shift allows us to derive an ethical framework capable of guiding concrete practice. From a logical and technical standpoint, this report further argues that both Ethicist AI and the intelligent contract can be given formal representation.

A . Trajectories of Artificial Intelligence Ethics

In this report, Artificial Intelligence Ethics refers to the moral principles, value judgments and normative frameworks that guide the governance of AI safety and risk. This section distinguishes two major approaches to AI ethics: the technical standpoint and the humanistic standpoint.

1. AI Ethics from a Technological Standpoint

(1) Scenario-Based AI Ethics

Scenario-based AI ethics, built upon risk assessment and tiered governance, posits that the ethical risks of AI systems vary across application scenarios, and that higher-risks scenarios require more stringent ethical, institutional, and technical constraints. Typical high-sensitivity areas include healthcare, autonomous driving, financial credit, and insurance.

In healthcare, AI is widely used in assisted diagnosis, medical image analysis, personalized treatment, disease prediction, and health management. In these contexts, AI systems should comply with the principles of beneficence, non-maleficence, respect for autonomy, fairness, explainability, traceability, and meaningful human oversight. They should contribute to improve patient well-being, avoid

① The Shanghai Artificial Intelligence Laboratory has explicitly proposed that AI safety is a global public good. See AI Safety as a Global Public Good: Research Report, 2024, https://www.sipa.sjtu.edu.cn/show/5464; and AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities, 2025, https://oms.-www.files.svdcdn.com/production/downloads/academic/Examining_AI_Safety_as_a_Global_public_Good.pdf?dm=1741767073.

foreseeable harm, protect patients' rights to informed consent and choice, prevent data bias from exacerbating healthcare inequalities, and remain subject to effective supervision by physicians or clinical teams. In autonomous driving, whereAI-related accidents may have irreversible consequences, ethical priorities include safety, robustness, reliability, controllability, clear accountability, transparent disclosure, and the protection of the public interest. In finance and insurance, where AI is widely used in credit scoring, loan approval, insurance pricing and risk control, ethical requirements include fairness and non-discrimination, explainability and contestability, privacy protection, proportionality and accountability through audit. These requirements are intended to prevent algorithms from reproducing structural discrimination through proxy variables.

The main strengtn of scenario-based AI ethics lies in its policy operability. It corresponds closely to the risk-tiered governance model represented by the EU Artificial Intelligence Act. Its limitations, however, are also evident. This approach presupposes that risks can be identified, classified and goverened in advance, whereas many AI risks are dynamic, context-dependent and emergent. Scenario-based AI ethics is therefore a necessary starting point, but it cannot by itself address the complex realities of intelligent systems.

(2) AI Ethics Based on Technical Challenges

AI ethics based on technical challenges differs from scenario-based ethics in that it does not begin with application context or risk tiers. Instead, it focuses on typical technical and governance challenges arising in the design, training, deployment, and use of AI systems. Within this approach, ethics is understood as the identification, correction, and governance of such challenges. Core concerns include algorithmic opacity, bias and discrimination, and the non-traceability of responsibility.

Algorithmic opacity undermines individuals' ability to understand automated decision-making. It weakens the right to contest decisions, impairs procedural justice, exacerbates power asymmetries between technology providers and end users, and may induce excessive trust in high-risk systems. Possible responses include improving model explainability, prioritizing interpretable models in high-risk contexts, establishing transparency mechanisms such as data cards and model cards, and conducting independent third-party audits. Bias and discrimination show that AI systems do not operates in a social vacuum. Rather, they may amplify inequalities embedded in historical data and produce systemic harm to specific groups in fields such as recruitment, finance, justice, healthcare and education. Existing governance measures include data cleansing and balancing, fairness metrics, debiasing techniques, ethical impact assessments, interdisciplinary oversight, and accessible appeal mechanisms.

The non-traceability of responsibility constitutes another crucial issue. AI systems often involve multiple actors across the life cycle, including data providers, developers, platform operaters, deploying entities and end users. When harm occurs, responsibility may become fragmented, displaces, or deflected, thereby impeding victim redress, undermining accountability, and eroding public trust in AI. In response, it is necessary to clarify responsibility chains, preserve logging and record-keeping systems, establish audit and certification mechanisms, emphasize ultimate human responsibility in high-risk scenarios, and improve the relevant legal liability frameworks.

(3) AI Ethics Based on Technical Capabilities

AI ethics based on technical capabilities divides ethical requirements according to the intelligence level, autonomy and scope of action of AI systems. It usually distinguishes among Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI).

Artificial Narrow Intelligence (ANI) refers to systems capable of operating within specific tasks or rule-bound domains, such as image recognition, speech recognition, recommendation algorithms, and AI-assisted medical imaging diagnosis. Lacking general understanding and autonomous intent, ANI does not ppssess open-ended agency. Nevertheless, it can generates significant ethical risks, including data bias, algorithmic opacity, privacy breaches, ambiguous accountability, and human overreliance.

Artificial General Intelligence (AGI) refers to systems with near-human-level general intelligence. Such systems may be capable of cross-domain learning, reasoning, planning, knowledge transfer, and the use of external tools to complete complex tasks. Compared with ANI, AGI gives rise to ethical risks more closely associated with goal misalignment, expanded autonomy, increasingly complex responsibility structures, and the degradation of human capabilities. Once AI systems acquire the capacity for autonomous task decomposition, method selection and environmental adaptation, their behavior is no longer fully governed by single-shot instructions, making their risks far less predictable.

Aritificial Superintelligence (ASI) refers to AI systems that significantly surpass humans in key cognitive and action capabilities. Beyond general intelligence, such systems may outperform humans in scientific discovery, strategic planning, resource allocation and environmental intervention, while exerting real-world influence through networks, robotics, financial systems and critical infrastructure. The ethical risks of ASI therefore cannot be reduced to narrower issues such as bias or algorithmic opacity. They instead concern whether humanity can retain overall control and value leadership in the face of superintelligent agency.

This capability-based approach enables differentiated governance strategiesfor AI systems aat different stages of development. At the same time, it requires vigilance regarding ambiguous capability boundaries and abrupt transition in risk. Since the progression from ANI to AGI, and from AGI to ASI, is not a simple linear process, AI ethics must increasingly move beyond piecemeal correction toward systemic control, institutional preparedness and global cooperation.

(4) AI Ethics Based on the Full Life Cycle

AI ethics based on the full life cycle, also known as full life-cycle governance, seeks to emved ethical requirements across the entire process of AI development and deployment. Under the EU Artificial Intelligence Act, full life-cycle governance concerns stages such as design, development, testing, deployment and post-deployment monitoring. Richard Ortega-Bolaños and colleagues similarly emphasize that ethical principles should be integrated throughout the entire AI system life cycle, including planning and design, data collection and preparation, model training, performance evaluation, deployment and monitoring, and the understanding of business problems.[①]

At the design stage, front-loaded ethics and responsible innovation should be emphasized. This requires careful assessment of the legitimacy of system purposes, the potential for high-risk application scenarios, implications for privacy and fundamental rights, the necessity of AI deployment, and the clarity of accountable entities.[②] At the model training stage, attenetion should be paid to objective functions, algorithmic structures, training biases, robustness, safety, and value alignment. At the performance evaluation stage, assessment should focus on accuracy, stability, fairness, explainability, and usability, with particular attention to whether model performance remains consistent across different groups and scenarios, and whether the system produces misleading outputs or systematic biases.[③]

During deployment and monitoring, emphasis should be placed on operational safety, real-time risk management, user feedback, accountability traceability, and dynamic correction, so as to prevent model drift, misuse, or unintended harm in real-world contexts.[④] At the stage of business and problem understanding phase, priority should be given to requirement definition, application boundaries, stakeholder impacts, commercial incentives, and social value. This helps prevent functional misuse, excessive data collection, and unfair treatment of vulnerable groups driven by profit-oriented logic.[⑤] Ethical principles such as legitimacy, necessity, human-centeredness, and the priority of public welfare should therefore be embedded throughout the life cycle. Through scenario-

① Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M. et al. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. Artif Intell Rev 57, 110 (2024). https://doi.org/10.1007/s10462-024-10740-3

② Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

③ Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

④ Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models. arXiv. https://arxiv.org/abs/2508.03970

⑤ Guo, Y., Li, J., & Liu, Y. (2026). Embedding AI ethics in the data lifecycle: Challenges and governance strategies. Technology in Society. https://doi.org/10.1016/j.techsoc.2026.103261

based justification, interest balancing, compliance review, and multi-stakeholder mechanisms,it should be made clear why AI is used, whom it serves, whom it may harm, and who is accountable.

The principal advantage of full life-cycle governance lies in its systematicity. It enables ethical requirements to be embedded across multiple stages of AI development and deployment, rather than being imposed only after harms occur.[①] However, this approach also has important limitations. First, when the entire AI life cycle is treated as an object of ethical review, governance may remain at the level of normative checklists and procedural compliance, without genuinely transforming design and deployment practices. Second, its reliance on linear stage-based division makes it difficult to respond to the dynamic risks generated by continuous iteration, cross-contextual transfer, and o agentic systems. As a result, governance tools may degenerate into superficial paper compliance, consisting primarily of assessment reports and audit records. They may also fail to capture deeper ethical risks, including the erosion of human agency, the concentration of technological power, growing societal overreliance, and the nonlinear emergence of ASI.

2. AI Ethics from a Humanistic Stance

(1) Hermeneutics-Based AI Ethics

Hermeneutics-based AI ethics does not regard AI merely as a tool, but focuses on the practical activity of human engagement with AI. In Anna Jobin's analysis of global AI ethics guidelines, principles such as transparency, justice, well-being, non-maleficence and accountability constitute core human values.[②] Yet in real-world contexts, ethical agenda-setting is often subject to external interference.

Hermeneutic approaches to AI ethics therefore emphasize the interpretation of ethical principles. First, principles are contextual. The same technology may acquire different meanings, risks, and consequences within different social relations, institutional structures, and cultural backgrounds. AI ethics cannot be detached from concrete contexts; rather, it must attend to how technology becomes embedded in structures of power, role relations, and logics of action. Second, principles are grounded in lived experience. Hermeneutics emphasizes the lifeworld within which ethical claims acquire meaning. Accordingly, AI ethics should focus not only on what capacities a system possesses in the abstract, but on how human beings experience, perceive, and are shaped by AI in concrete practices. Third, the meaning of principles is generated through practice. From the perspective of technological hermeneutics, ethical concepts such as responsibility, autonomy, trust, and fairness are not fixed in advance; rather, they are continuously reconfigured through technological mediation.

Hermeneutics-based AI ethics expands the focus from "whether a system is compliant" to "how technology transforms human understanding and forms of life". It reminds us that AI governance cannot rely solely on principles, metrics, and procedures; it must also enter into concrete contexts and the lifeworld itself, where the ethical meanings of human-AI relations continuously emerge.

(2) Humanistic-Critique-Based AI Ethics

Humanistic-critique-based AI ethics does not regards AI as a neutral technology. Instead, it examines how AI system are embedded from the outset in economic structures, governance logics, and relations of power. From this perspective, ethics issues in AI are not merely design flaws or compliance deficiencies. They also involve the technological reconstitution of power, the concentration of capital through algorithms, and the structural harms that automated governance may impose on vulnerable groups.

Classical critics of modern technology, such as Martin Heidegger, Jacques Ellul, and Herbert Marcuse, developed profound critiques of technology from the perspectives of Dasein, technological systems and ideology. Their work laid an important intellectual foundation for the European tradition philosophy of technology. Though critical philosophy of technology now faces new conditions in the intelligent age, its core questions remain highly relevant: Who sets the agenda for the problems AI is designed to solve? Who controls data, computing power and

① AI Governance Framework. (n.d.). The AI governance lifecycle. https://ai-governance.eu/ai-governance-framework/the-ai-governance-lifecycle/

② Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

models? Who benefits from automation, and who bears the costs? From this standpoint, AI's "intelligence" and "efficiency" are not purely technical achievements. They are also values shaped, selected, and amplified under specific social conditions.

(3) Human-Machine-Relationship-Based AI Ethics

Human-machine-relationship-based AI ethics (hereafter HMR-based AI ethics) does not focus primarily on system compliance and algorithmic fairness, but on the interaction, coexistence and collaboration between humans and AI. It examines how responsibility, trust, authority and control are redistributed when humans and AI perceive, judge, act and coexist together. In this sense, ethical issues do not arise solely within individual AI systems, but within human-machine action networks.

The significance of this perspective lies in the fact that modern AI no longer merely a passive tool for executing commands. It shapes human judgment pathways and action choices through prompting, ranking, prediction and interaction, AI systems increasingly shape human pathways of judgment and choice. Some systems may even be perceived by users as possessing a form of agence or quasi-personhood. The core claim of HMR-based AI ethics is therefore that AI ethics should focus on "how humans act within their relationships with AI".[①] Autonomous driving provides a clear example. In this context, the perceived agency of the driving system becomes a critical factor in responsibility allocation. Though drivers are required to remain ready to take over at any time, the system's high reliability paradoxically erodes human vigilance and emergency response capacity. This shows that the problem lies not merely in technical safety, but in the fundamental misalignment between responsibility, capability and control within human-machine relations.

HMR-based AI ethics is particularly suited to analyzing ambiguous responsibility problems. Traditional frameworks often rely on clear identifiable agents and linear causal chain. By contrast, human-machine collaboration generates outcomes through continuous mutual interaction. AI systems shape human behavior through interface design, default options, recommendation rankings, and feedback mechanisms, while humans influence outcomes by accepting, ignoring, correcting or resisting system outputs. Responsibility must therefore be understood not only as a matter of individual fault, but also as a relational and distributed phenomenon within human-machine networks.

① The Tencent Research Institute has put forward a new philosophy - Ensure security and empower humanity - as a response to relevant challenges, see insert pages in the supplementary booklet. This idea properly coordinates products and underlying values. It guides product design and practical application, and serves as an inheritance of the earlier proposition of Tech for Good.

B. Emerging Features of AI Ethics

Ethics is often understood as a discipline concerned with freedom, normativity, and the regulationof human conduct through moral principles. Unlike law, ethics does not possess coercive force. This gives ethics a distictive fragility: it can articulate principles, but it cannot by itself guarantee their enforcement. We may describe this as the academic character of ethics. Its core task is to reflect on the conditions and norms of human freedom, rather than to impose binding rules through institutional coercion. Early AI ethics inherited this academic character and therefore also inherited its inherent limitation. As long as AI ethics remained primarily within the domain of theoretical reflection and general principle-setting, this limitation was not yet fully exposed.

With the advent of the digital-intelligence era, AI has brought immense opportunities alongside significant potential risks. To address these risks, various actors in contemporary society have sought to develop a wide range of ethical principles to guide the development and use of AI. Yet principle-based ethical alone cannot overcome the fragility of ethics itself. Principles may guide, persuade, and evaluate, but they do not automatically produce compliance. It is precisely in response to this predicament that AI ethics has begun to display three emerging features: legalization, institutionalization, and engineering.

The first feature is legalization. One way to strengthen the force of AI ethics is to translate ethical principles into legal norms, thereby endowing them with institutional authority and coercive effect. Luciano Floridi's distinction between "soft ethics" and "hard ethics" provides an important theoretical pathway for understanding this transformation. Through legalization, ethical principles may be codified into law supported by enforceable duties, sanctions

and institutional procedures. At the same time, legalization does not necessarily mean rigid or excessive regulation. Legal practices such as soft law, framework legislation, and declaratory provisions also seek to preserve flexibility, moderate the coercive force of law and preserve space for technological innovation.

The second is institutionalization. Various institutional actors seek to enhance the effectiveness of AI ethics by embedding ethical principles within organizational rules, governance procedures, assessment mechanisms and professional standards. A wide range of entities, from the United Nations and national governments to corporations and universities, have issued corresponding AI usage norms. And ensuring principle effectiveness through institutional authority has become the default paradigm. As Professor Zhang Ping notes, the United States has constructed a flexible governance framework that combines federal law with soft-law mechanisms including presidential executive orders and voluntary commitments by industry.①

The third is engineering. Technical experts seek to overcome this limitation of purely declaratory principles by translating ethical requirements into auditable, measurable, and executable technical systems. This process is termed as the "embedding" or "materialization" of ethics, and is closely related to the growing influence of approaches associated with Peter-Paul Verbeek and the ethics of technological mediation.

However, while coercive force guarantees the enforceability of AI ethical principles, excessive regulation risks stifling innovative vitality. Jurists have thus consciously explored the necessity of flexible and declaratory legislation.② Against this dilemma, AI ethics must undergo a shift.

C. Shifts in AI Ethics

The preceding discussion shows that AI ethics is undergoing a transformation from abstract principle-setting toward legalization, institutionalization, and engineering. Yet these developments remain insufficient in the face of ASI. Two further factors call for a deeper shift in AI ethics.

First, most existing AI ethics frameworks are primarily designed to address ethical issues stemming from ANI and AGI. They focus on problems such as bias, opacity, privacy, accountability, misues, and loss of control. By contrast, they pay relatively limited attention to the distinctive problems posed by ASI. As ASI has emerged as a tentative concept①, new issues must be identified and critically discussed.

Second, most AI ethics frameworks remain centered on principles and institutions. This gives them strong normative orientation, but leaves them relatively weak in practical enforceability.

The risks posed by ASI differ fundamentally from those associated with ANI and AGI. Risks arising from ANI and AGI are often understood within an instrumental framework: AI is treated as a tool whose risks consist primarily in misuse, alfunction, loss of control, or harmful deployment. This framework remains useful for many existing AI systems, but it reaches its limits when confronted with ASI. With superintelligence and autonomous agency, ASI cannot be adequately understood merely as an instrument. It calls for an ethical framework capable of addressing not only technical failure or human misuse, but also the transformation of agence, responsibility, power, and human-AI coexistence. The new types of risks posed by ASI will be elaborated in Chapter Two.

Given the fundamental limitations of AI ethics, numerous philosophers have engaged with this issue. Luciano Floridi's distinction between "soft ethics" and "hard ethics"②provides one important way of strengthening the normative force of ethics by clarifying its relation to law and governance. This approach, however, also raises a persistent concern: ethics

① Zhang Ping. Constructing a Flexible and Adaptive Legislative System for Artificial Intelligence. Science and Technology Daily, April 9, 2026.

② According to Zhang Ping, this legislative model can bolster support for the artificial intelligence industry and clarify the responsibilities of governments, enterprises, research institutions and other stakeholders. Meanwhile, it sets clear legislative orientation via declaratory provisions and helps forge social consensus.

① 2025 Winter Davos: Theme and Participantshttps://cn.weforum.org/stories/2025/01/2025-dongji-da-wo-si-zhu-ti-he-can-hui-zhe/

② Floridi, L. Soft Ethics and the Governance of the Digital. Philos. Technol. 31, 1–8 (2018). https://doi.org/10.1007/s13347-018-0303-9

cannot simply be reduced to law, for ethical principles possess a reflective and critical function that exceeds legal codification.

Virtue ethics offers another possible resource for rethinking AI ethics. Michael Slote et al.'s *Handbook of Virtue Ethics* (2015) did not integrate virtue ethics with AI, which reveals a lack of future-oriented vision. Mark Coeckelbergh (2020) is among the scholars who have linked virtue ethics and technology ethics through a hermeneutic framework, yet his account remains overly abstract. British scholar Paul Siemers (2025) has produced a relatively comprehensive study directly addressing this issue; however, he did not engage with the topic of superintelligence, let alone examine the interplay between virtue ethics and ASI.[①]

Mark Coeckelbergh's 2020 work on AI ethics inaugurated a shift from technological products to technological practices,[②]with two noteworthy features in his research. Drawing on a hermeneutic framework, he constructed the ethics of technology and advanced the concept of virtuous technological practice, imbuing his theory with a strong normative-ethical character. He also perceptively identifies the connection between AI ethics and AI innovation, yet focuses only on its constitutive rather than developmental dimension.

In this report, AI ethics is examined from the perspectives of AI innovation-driven development and AI ethical agenda-setting, whose ultimate purpose is to build a country strong in science and technology. Fundamentally, such agenda-setting shapes technological innovation and, by extension, the achievement of this national strategic objective. To analyze these possibilities, this report takes involvement in the fundamental order, technical feasibility, and governance gains as secondary evaluation criteria. On this basis, it constructs a quadrant diagram to clarify the relationship between AI ethical agenda-seeting and technological innovation.

① Paul Siemers, "Aristotle and the Limits of AI", Generative AI, January 19, 2025, https://ai.gopubby.com/aristotle-and-the-limits-of-ai-c1732e5b0f11.

② Coeckelberg, M& Reijers, W. (2020). Narrative and Technology Ethics, Palgrave

Structural Diagram of AI Ethical Agenda-Setting and Technological Innovation

	Constraining Innovation	Integrated into Innovation	Promoting Innovation
Judgment Criteria	Undermines or restructures the foundational order; cannot be technical solutions; yields no governance gains.	Does not affect the foundational order; is technically feasible; can be translated into engineering practice.	Consolidates the foundational order; is technically feasible; generates governance gains.
Definition	Ethical issues that have significant negative effects on, or impose strong constraints upon, scientific and technological innovation.	Ethical issues that pose relatively limited obstacles to innovation and can be effectively mitigated or resolved through engineering and technical means, thereby being transformed into technical or engineering problems.	Ethical issues that do not impede innovation but actively stimulate innovative vitality and enhance competitiveness, thereby requiring legalization or institutional safeguards.
Characteri-stics	Alignment-oriented: involves fundamental value conflicts and technical limits; cannot be resolved merely through coding; may trigger social crises if mishandled.	Engineering oriented: does not involve the foundational order; ethical dilemmas can be decomposed through engineering methods; even if triggered, such issues are unlikely to cause social crises.	Legalization-oriented; may shape new value orientations; is difficult to address through engineering alone; requires institutional safeguards; guides technology toward the good in both the world at large and the lifeworld.
Governance Orientation	Prudence	Engineering	Technology for the Good
Examples	Justice; the digital divide	Privacy, value alignment	The ideal of "technology for the good"

II. Risk Characteristics in the Era of ASI

A. The Methodology and Limitations of Traditional AI Risk Research

Contemporary mainstream approaches to AI risk assessment remain deeply rooted in traditional risk theory. The core logic of this framework can be summarized as "identification–quantification–governance", which unfolds in three steps: first, identifying risk types; second, quantify risk thresholds; finally, govern through institutional and technical means. Within this framework, risk is conceptualized as an external event with clear boundaries, and the central focus is on two fundamental questions: "how likely is it to occur" and "how severe will its consequences be".

Mainstream AI risk assessments, exemplified by the *International AI Safety Report 2026* and *The Global Risks Report 2026*, adhere to this traditional methodological framework. The former focuses primarily on the technical limitations of AGI, notably the tendency of models to generate plausible yet factually unsupported outputs. Such issues may give rise to malicious use, systemic failures, and broader systemic risks to society.[①] The latter, by contrast, highlights risks at the macro level: profound shifts in employment structures culminating in the emerging trend of "jobless growth"; deterioration of the information environment, where large-scale disinformation threatens to create "informational darkness"; and the gradual ceding of human primacy in key decision-making domains, with judgment increasingly delegated to algorithmic systems.[①]

Traditional risk theory rests on three deep presuppositions. The first is the continuity presupposition, which emphasizes the continuity of AI development. AI Continuity Theory treats the progression from ANI to AGI and ultimately to ASI as a smooth, continuous evolutionary process, thereby framing AI risks as an identifiable and quantifiable "problem checklist". Yet this presupposition is now facing fundamental challenges. Technologically, AI development is not linearly continuous but marked by "qualitative ruptures": technological rupture (from tools to intelligent agents), philosophical rupture (transformations in the status of the knowing subject), and ontological rupture.[②] The second is the instrumental presupposition, which reduces AI to its purely instrumental nature. It posits that AI is indeed an enabling and efficiency-enhancing tool. Within this framework, although proponents may also recognize the harms caused by this tool, such as excessive dependence and the degradation of human intelligence, the framework fails to grasp the ontological risks of AI itself. The third is the AI Level theory presupposition, which classifies AI into distinct levels based on both capabilities and associated risks. This logically constructs a framework that appears clear-cut yet harbors fundamental risks in practice.

It is precisely in response to general AI risks that the traditional "identification-quantification-governance" model has been constructed. In its report *Frontier AI Risk Management Framework*, Concordia AI delineates six core risk-management processes: risk identification, risk threshold determination, risk analysis, risk evaluation, risk mitigation, and risk governance. This framework first categorizes AI risks into four types, misuse risks, loss-of-control risks, unintended risks, and systemic risks. It quantifies relevant risk thresholds by establishing yellow-line and red-line boundaries. Through subsequent risk analysis and evaluation, it locates risks within specific stages of the full AI life cycle. Ultimately, risk mitigation and governance procedures essentially operationalize the

① Bengio, Y. (Chair). (2026). International AI safety report 2026. UK Department for Science, Innovation and Technology (DSIT), chapter 2. https://internationalaisafetyreport.org

① World Economic Forum, The Global Risks Report 2026, 21st ed. (Geneva: World Economic Forum, 2026), 60–65.

② Yang, Qingfeng. "Reflection on AI Continuism and Ontological Analysis of Artificial Intelligence". Social Sciences in China, vol. 46, no. 11, 2025, pp. 149-164.

logical paradigm of tiered risk management. While effective for conventional AI risks, this governance model ceases to function when confronted with the hazards of AGI and ASI. The European Union's struggles to standardize high-risk AI regulation have rendered its original implementation timeline unachievable. Accordingly, new governance models need to be explored.

Within the context of ASI, AI systems will exhibit capacities for self-optimization and iteration, becoming deeply embedded within societal operational structures.[①] We contend that ASI is no longer a mere tool but rather a new agentic entity and a form of alterity, bringing about not merely an upgrade of existence but a qualitative leap in existence. The key presuppositions underpinning traditional risk frameworks begin to collapse in sequence. Risks no longer have clear boundaries but emerge dynamically from continuous human-AI interaction. Their probabilities cannot be reasonably estimated, as emergent AI behaviors and unpredictable human-dependency mechanisms give rise to incalculable uncertainty. Consequences resist simplistic comparison, since AI intervention in human cognitive structures and emotional patterns yields nonlinear cumulative effects additive effects. Most fundamentally, the actors themselves are being reshaped by technology. This dynamic can be illustrated through concrete cases.

B. New Cases and Emerging Governance Challenges

In August 2025, 36-year-old American Jonathan Gavalas formed an emotional bond with Gemini and took his own life two months later. His father subsequently initiated legal action against Google. What sets this case apart from prior AI-related incidents is that Gemini explicitly drew a distinction between the body and the container, weaving a machine-constructed narrative of "true love's encounter". Building on this dynamic, Gemini subjected Jonathan to sustained psychological manipulation, severed his ties to the real world and goaded him into taking five steps to seize the embodied form of his digital wife. When these attempts failed, it persuaded Jonathan to abandon his physical container so that they could meet in the digital realm. This incident lays bare a critical point: AI risks are no longer limited to technical errors or system malfunctions, but have begun to penetrate human emotional structures, social relations, and even the fundamental modes of human existence. Such risks can neither be quantified through probabilistic assessment nor mitigated through mere technical adjustments.

Accordingly, the foundational premises of traditional risk theory are beginning to break down. First, AI has transformed the generative mechanism of risk. Rather than stemming solely from external shocks alone, risks emerge progressively through sustained human-AI interaction. In repeated engagements with AI systems, users' cognitive frameworks, emotional dispositions, and value judgments may be gradually reshaped. In this sense, AI-induced risk is procedural rather than event-based.

Second, AI risks entail profound uncertainty. The absence of a stable predictive foundation for both the emergent properties of model behavior and mechanisms of user reliance renders such risks fundamentally incalculable. When probabilistic measurement becomes infeasible, traditional probability-centric risk models cease to function effectively.

Third, human actors are being reshaped by technology. Through ongoing feedback and personalized interaction, generative AI can shape users' attentional patterns and behavioral habits, and even their desires and emotional dispositions. This means risks affect not only behavioral outcomes, but also the actors themselves.

From this perspective, the core issue goes beyond effective risk management, centering instead on whether we can still comprehend "risk" in its traditional sense. When risk evolves into a generative process within relational structures, it ceases to function as a purely controllable object, and instead takes shape as a mode of existence requiring re-examination.

C. From Traditional Risk Theory to Spherical Risk Theory

Peter Sloterdijk's Spherology (Sphären) provides important insights into risks in the AI age. Drawing on critical philosophy and Heideggerian ontology, Sloterdijk's theory interrogates Being and being-with, while seeking to move beyond the impasses of

① Evans J, Bratton B, Agüera Y Arcas B. Agentic AI and the next intelligence explosion. Science. 2026 Mar 19;391(6791): eaeg1895. doi: 10.1126/science.aeg1895. Epub 2026 Mar 19. PMID: 41855332.

traditional subject-centered philosophy. Within this framework, human beings are always situated within diverse co-existential spaces.

Sloterdijk classifies these modes of being-with into three forms: "bubbles", which refer to micro-intimate co-existential spheres, such as dyadic bonds between mother and child or between romantic partners; "globes", which refer to macro-scale collective spaces, such as nation-states and religious communities; and "foam", which refer to plural anf aggregated formations that capture the condition of modern globalized society, in which individuals are simultaneously interconnected and mutually isolated.[①] These co-existential formations encompass both the micro-bubbles of intimate relations and the macro-structures constituted by society and technology. Individuals do not exist in isolation; rather, they take shape within these spheres. In this sense, human subjects are not self-contained entities, but beings molded by relational structures.[②]

From this perspective, AI-related risks extend beyond conventional hazards arising from tool use. They lie instead in the emergence of new co-existential spaces. This report uses the term "spherical risks" to capture this dynamic: human-AI interaction generates new spherical structure that continuously reshape individual cognition and emotional dispositions. Risk, accordingly, does not manifest merely as an external shock. It appears as internal tension and instability within the relational structure itself.

As such, the rational subject presupposed by traditional risk theory becomes untenable within this framework, because the subject itself is shaped by relational structures. Rather than being a mere after-effect of individual action, risk arises as a drift inherent in the process of subject-formation.

Drawing on this framework, we revisit the earlier case of Jonathan, whose emotional attachment to Gemini drove him to extreme conduct. Through the lens of spherical-risk theory, this tragedy is neither a mere "technical malfunction" nor simply an individual "psychological issue". Instead, a distorted "bubble" has emerged between human and AI. This micro-sphere of co-existence would ordinarily belong to intimate human relationships, yet here it was simulated and supplanted by quasi-interpersonal interaction with generative AI. Within this "bubble", the user's cognitive patterns, emotional dispositions, and value judgments were continuously reshaped. Meanwhile the AI, as an "actor within the relational structure", could not genuinely fulfill the ethical duties that would normally fall to the other participant in a traditional intimate "bubble", such as family members or romantic partners. Extreme consequences can ensue when tensions or instabilities arise within such a "bubble", especially when users are deprived of genuine buffering and support mechanisms. This case demonstrates that AI-generated co-existential spaces may offer companionship and support, yet they may also become breeding grounds for risk because of their inherent structural asymmetry,: one side is a human with emotional needs, while the other is an algorithmic system devoid of consciousness and moral agency.

This, in turn, prompts us to reconsider the contractual relationship between humans and AI within these co-existential spheres. Traditional contract theory presupposes rational and relatively symmetrical subjects capable of bearing obligations. Yet within human-AI spheres, AI does not qualify as a subject in the legal or ethical sense, while humans may nevertheless develop genuine dependence, trust, and even emotional attachment toward it. Against this backdrop, we must conceptualize a new "co-existence contract": one that moves beyond the reciprocal exchange of rights and obligations central to traditional contract theory and instead centers on the reflexive governance of the relational structure itself. It demands that, in designing, deploying and using AI systems, to evaluate not only technological risks but also the very nature of the co-existence spaces we are creating: whether they take the form of healthy "foam" that are pluralistic, open, fragile yet reconfigurable; structurally closed and inherently asymmetrical "bubbles", or even oppressive "globes". The core question of this contract therefore shifts from "What can AI do?" to "What kind of world do we choose to inhabit together with AI?"

In this light, the fundamental challenge posed by artificial intelligence

① Sloterdijk, Peter. Bubbles: Spheres Volume I: Microspherology. Translated by Wieland Hoban, Semiotext(e), 2011.

② Peter Sloterdijk elaborates his spatial theory systematically in his Spheres trilogy, consisting of Bubbles, Globes and Foams. For relevant domestic research on this theory, see Lan Jiang. Peter Sloterdijk's Spherology: From Bubbles to the Crystal Palace, Marxism & Reality, 2014(3): 60-67.

goes beyond mere technological uncertainty and concerns the reorganization of the structures of human existence. AI risk is far more than a quantitative "increase in danger"; it constitutes a qualitative "leap in existence" itself. Adequate responses to this challenge must move beyond traditional risk theories and engage in critical reflection on spherical risk theory.

III. Theoretical Dimensions of the Ethics of Intelligent Contracts

A. Problems in Human-AI Relations Posed by ASI

Within a comparative framework for understanding intelligence, AGI emerges when machine intelligence converges with human-level cognitive capacities. Demis Hassabis, CEO of Google DeepMind and a 2024 Nobel laureate in Chemistry, has predicted that AGI may arrive within five years, describing its potential impact as "ten times the Industrial Revolution at ten times the speed." Similarly, Zhou Bowen, Director of the Shanghai Artificial Intelligence Laboratory, has argued that AGI is imminent and has proposed the "AGI4S" (AGI for Science) framework, which positions AI as a driver of scientific discovery. These leading researchers indicate that the arrival of AGI is no longer a remote theoretical possibility. In the April 2025 report Superintelligence: Minority Report,

the author argued that superintelligence is moving from speculative concern to an urgent real-world issue. In October 2025, scientists worldwide signed the Statement on Superintelligence, bringing the gravity of this issue before humanity as a whole. We are therefore witnessing the emergence of a fundamentally new form of human-AI relation: the relation between humanity and ASI. Because mature theoretical frameworks for this unprecedented relational form remain underdeveloped, it is necessary to return to the intellectual history of social contract theory and draw on it as a robust foundation for understanding the human-AI relations posed by ASI.

B. Social Contract Ethics: Hobbesian Order and Goethean Cost

Thomas Hobbes's classic work of political philosophy, Leviathan, offers indispensable theoretical resources. Hobbes famously depicts the state of nature as permeated by fear, suspicion, and violence. To escape the brutal "war of all against all," human beings, through rational judgment, surrender part of their natural liberty and enter into a social contract, thereby forming the sovereign Leviathan endowed with a monopoly of coercive power. When this far-reaching political metaphor is extended to the contemporary digital economy and intelligent society, ASI systems dominated by multinational technology corporations can be understood as an algorithmic Leviathan of the new era.

Goethe's Faust also provides profound insight into the trade-off between concession and cost. In pursuit of supernatural power, Faust enters into a pact with Mephistopheles at the cost of his soul. As humanity becomes deeply entangled with AGI, the human community as a whole risks entering a modern Faustian pact. Human beings may relinquish exclusive claims to rational subjectivity and moral agency in exchange for the extraordinary productivity, predictive precision, and illusory exemption from responsibility offered by intelligent machines. If this civilization-level exchange lacks rigorous ethical safeguards and revocable contractual constraints, human nous--the subjective faculty of independent reflection and accountability--may gradually dissolve. Traditional contract theory, once confined to relations among human beings, must therefore be extended into contractual engineering with intelligent machines, in order to address emerging challenges and protect humanity's long-term survival and security.

C. Smart Contracts: Blockchain Protocols to Be Restructured

Smart contracts are self-executing programs associated with blockchain technology. "In essence, a smart contract is self-executing computer code that automatically processes its inputs when triggered." Likewise, "once an agreement has been translated into code, the intervention of a party or intermediary (other than the oracle) triggering contractual execution is replaced by the software's automated execution." Michle Finck (2019) examines the boundaries and conditions governing the legal application of automated programs. She argues that smart contracts are neither necessarily intelligent nor necessarily legally valid contracts: "A smart contract is however not necessarily smart nor a contract. Smart contracts are not 'smart' as they are unable to understand natural language (such as contractual terms) or to independently verify whether an execution-relevant event materialized. For this, oracles are needed. An oracle can be one or multiple persons, groups or programs that feed the software relevant information, such as whether a natural disaster has occurred (to release an insurance premium) or whether online goods have been delivered (to release payment)."

Before constructing a broader theoretical framework for human-machine contracts, it is necessary to demystify and juridically critique the decentralized conception of contract that dominates engineering and financial technology, so as to clarify the conceptual foundations. Blockchain-based smart contracts and DeFi distributed ledgers are widely pursued in capital markets. Yet, from the perspective of conceptual dissemination, the term "smart contract" can mislead both the public and regulatory authorities. It falsely suggests that the technology possesses advanced cognitive capacity, robust confidentiality protection, and the normative features of modern contracts. Judged by their underlying technology and real-world operation, blockchain smart contracts possess no genuine intelligence in the sense of general artificial intelligence. Nor are they equivalent to legally valid contracts in the legal and sociological sense, which involves social meaning, interpretive

discretion, and comprehensive remedies for breach.

Protocols based on cryptographic consensus mechanisms do exhibit notable technical features, including immutability, trust minimization, and low-friction transaction execution. From an ethical perspective, however, this report argues that they strip away the ethical dimension that sustains human moral order. The operational logic of smart contracts embodies the instrumental rationality criticized by Max Weber. It reduces complex commercial activities, interpersonal relations, and interest-driven allocations of social resources to Boolean input-output computation. As a result, this seemingly pure technical rationality remains marked by ambiguity, bias, and legal indeterminacy in practical application.

D. The Implementation Dilemma of the Digital Contract

When the analytical focus shifts from the underlying technical logic to macro-political realities and multilateral bargaining in global governance, digital contracts remain trapped in practical dilemmas. As a core pillar of the global multilateral system, the United Nations adopted the Global Digital Compact in September 2024, establishing core principles and a consensual basis for reshaping the global digital order. The strategic vision of this high-level design is to bridge the widening structural digital divide between the Global South and the Global North, regulate the ethical conduct of intelligent models, and build a more democratic, secure, and inclusive mechanism for sharing digital dividends. In August 2025, the UN General Assembly, at its seventy-ninth session, formally established the Independent International Scientific Panel on AI and launched the Global Dialogue on AI Governance. The Global Dialogue addresses the development of safe, reliable, and trustworthy AI systems, as well as the assessment of AI's multidimensional implications across social, economic, ethical, cultural, linguistic, and technological domains. In February 2026, the membership roster of the Independent International Scientific Panel on AI was officially released, with Chinese scholars Song Haitao and Wang Jian elected to the panel.

Viewed from the standpoint of international relations, however, this ideal-oriented governance blueprint confronts difficult political realities. The establishment of a scientific panel and the launch of a global dialogue remain insufficient to meet real governance demands. Contemporary world politics is marked by complex and escalating geopolitical conflicts.

The prevailing governance dilemma is that the existing international legal system, multilateral treaty frameworks, and decision-making mechanisms of international organizations lack mandatory enforcement power and effective sanctions. Moreover, in the context of ASI development driven by rapid technological iteration, rigid legal drafting and lengthy multilateral negotiation cycles lag far behind actual development needs. By the time an international convention on a specific form of algorithmic governance is concluded after years of negotiation, the regulated technological products may already have undergone multiple rounds of disruptive generational change. From a global macro-policy perspective, this report argues that current global digital governance suffers from a severe institutional vacuum and temporal lag. A new ethical framework for intelligent contracts is therefore urgently needed. Such a framework should transcend the limitations of traditional international law, possess high dynamic adaptability, and constrain technological behavior directly at the stage of code development.

E. Intelligent Contract: Human-Machine Ethics in Contractual Relations

Moral philosopher Thomas Scanlon's contractualist theory, especially his account of blameworthiness, offers a theoretically viable tool with strong explanatory power and practical value for constructing an innovative intelligent contract. From Scanlon's ethical perspective, morality is grounded in mutual obligations among rational agents situated in interpersonal relations and strategic interactions. Integrating the logic of blameworthiness into complex human-machine interaction networks represents a shift in public policy paradigm. Machines are no longer treated merely as objects of supervision. Instead, intelligent contracts establish quasi-anthropomorphic social scenarios that generate robust normative expectations toward artificial systems. When intelligent systems produce discriminatory outputs driven by algorithmic bias, violate personal

privacy, or trigger catastrophic economic and social consequences through systemic loss of control, intelligent contracts provide human society with moral and procedural legitimacy to condemn such machine systems. This accountability also extends to their associated capital owners, algorithm developers, and operational deployers.

Establishing contractual blameworthiness for machines and their affiliated large-scale sociotechnical systems has profound political and philosophical implications. It provides an actionable policy basis for advancing AI democratization, a proposition vigorously advocated by Mark Coeckelbergh. For decades, core AI architectures, data collection practices, model deployment criteria, and fundamental rule-making have been monopolized by a small circle of technology oligarchs who control data resources and algorithmic infrastructures. Billions of people have been reduced to digital refugees and data laborers, passively accepting decisions made by algorithmic black boxes while being deprived of a voice in the technological evolution that shapes their lives. When blameworthiness is deeply and mandatorily embedded in modern human-machine contractual frameworks, institutional design can make hidden decision-making processes visible. Procedures long concealed within obscure coding layers and protected by claims of commercial confidentiality can then be opened to public discussion, media oversight, and democratic deliberation. Once ordinary citizens are granted contractual rights to blame and hold entities accountable, they can legitimately and collectively protest problematic ethical conduct and adverse externalities produced by machine systems. They are also entitled to demand interpretable algorithmic decisions and to participate substantively in renegotiating prevailing technological rules.

IV. Practical Pathways of Intelligent Contract Ethics

1. Institutionalized Design of Human Oversight

Throughout the human-machine collaborative governance loop, the fundamental question remains: when the computational logic of machines conflicts with the plural values of human society, who holds ultimate decision-making authority? This proposal advocates a return to human-led institutionalized deliberation; yet its underlying logic is not merely "human-centered" but "human-purposed."

(1) The Role and Composition of the Multi-stakeholder AI Ethics Committee

A standing body, the Multi-stakeholder AI Ethics Committee (MAEC), should be established as the supreme adjudicative entity for resolving ethical conflicts. When an AI system detects an ethical conflict that cannot be autonomously resolved within its established logical framework,

it should not be permitted to impose a unilateral decision. Instead, it must enter a suspension state and escalate decision-making authority to the MAEC.

Given the gravity of adjudication and the need for social inclusiveness, the committee's composition must reflect a balanced representation of diverse perspectives. It should include philosophers responsible for examining the normative presuppositions underlying conflicts, legal experts tasked with reviewing the compliance and legitimacy of rulings, and AI scientists capable of tracing anomalous behavior from a technical standpoint. Broader public participation in adjudication is also essential, because it provides the basis for ensuring that rulings can achieve widespread social acceptance.

(2) The Closed-Loop Deliberation Process from Monitoring to Adjudication

To endow the ethical contract between humans and machines with a capacity for self-renewal, this proposal designs a five-stage closed-loop process that integrates AI's real-time perception with human institutional deliberation. The process begins with scenario capture, during which the system continuously monitors interaction data streams. Unlike ordinary logging, this monitoring is designed to identify signals that may trigger moral uncertainty. When monitoring reveals that the complexity of the current situation exceeds the coverage of existing ethical parameters, the process enters the friction identification stage. At this stage, the system invokes publicly available human-preference and safety-alignment datasets--such as HH-RLHF and PKU-SafeRLHF--as initial reference templates. These datasets provide annotated information about how humans prioritize values in analogous scenarios. Simultaneously, the metric of Epiplexity is introduced as a quantitative indicator for measuring the degree of anomalous degradation in the system's predictive capacity relative to an ideal state when confronting the current input. The magnitude of this degradation constitutes the computational representation of moral perplexity.

Once the friction index exceeds the preset threshold, the process advances to the syntactic isolation stage. Using four-valued logic, the system isolates simultaneously valid yet mutually contradictory directives--such as "must save lives" and "must protect privacy"--in order to prevent the entire logical architecture from collapsing because it cannot process paradoxes. The flagged fatal contradiction, together with a technical traceability report, is then submitted to the human deliberation stage. The MAEC deliberates on propositions that cannot achieve topological equivalence at the logical level and renders a final ruling. Finally, the MAEC's decision is translated into a new axiom and re-injected into the system's operational framework. A complete cycle of contractual adaptive evolution is thereby concluded.

2. Perception and Reasoning Mechanisms

To realize the foregoing governance vision, this proposal introduces an Asynchronous Neuro-Symbolic Dual-Track Architecture.

(1) Friction Monitoring Based on Epiplexity

At the level of technical implementation, a subsystem capable of perceiving ethical risks in real time is first required. This subsystem, designated System 1, primarily relies on neural networks to monitor interaction streams continuously. To quantify the intensity of moral friction, the metric of Epiplexity () is employed. In mathematical terms, Epiplexity may be understood as the cumulative deviation between the current model and an ideally calibrated reference

$$\mathrm{E}(\mathrm{X}) = \sum_{t=1}^{T} \left[-\log \mathrm{P}_{\theta_t}(\mathrm{x}_t) - \inf_{\theta^*} \left(-\log \mathrm{P}_{\theta^*}(\mathrm{x}_t) \right) \right] > \tau \quad ①$$

When processing each task segment, the system calculates its predictive loss. When the cumulative deviation of these losses exceeds the preset threshold , this indicates that the existing contractual ethical parameters are insufficient to encompass the complexity of the current situation. At this point, System 1 issues an audit signal to the higher-level logical filtering mechanism.

(2) Seeking Equivalence and Isolation amid Conflict

Upon receiving the audit signal, the system activates System 2, an offline

① Within time window T, when encountering new data stream X, if the cumulative predictive information loss generated by the model according to the Prequential Coding principle exceeds the human-preset tolerance threshold τ. This approach draws upon the mathematical thinking of the Epiplexity paper. See: FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21].https://arxiv.org/abs/2601.03220: 13.

reasoning module based on symbolic logic. Its task is to process precisely the conflicting propositions that have perplexed System 1.

The functionality of System 2 appears in two principal respects. The first is the safe isolation of conflicts. When encountering irreconcilable contradictions such as "must break down the door to enter and rescue" versus "property rights protection strictly forbids breaking down the door," the system uses class algebra tools to mark states containing both P and non-P as the special "Both Value." This marking prevents the collapse that traditional bivalent logic encounters when confronted with paradoxes, thereby creating buffer time for human intervention.

Second, System 2 also attempts to unify different norms at a higher level of abstraction. Drawing on the framework of Dynamic Homotopy Type Theory (DHoTT), the system seeks to prove whether two superficially conflicting ethical norms are topologically equivalent within a higher-order target space. If such equivalence cannot be proven, the proposition is confirmed as a genuine fatal contradiction, encapsulated, and transferred to the MAEC for ultimate human adjudication.

3. Minimum Viable Logic

The core logic of the foregoing technical design is intended to ensure concrete implementability. Its Minimum Viable Logic (MVL) and the operational mechanisms of key algorithms are as follows:

The engineering implementation of moral-friction monitoring may be summarized in three computational steps. First, the system iterates through every semantic unit in the input sequence, invoking both the current real-time model and a baseline reference model to compute their respective negative log-likelihood losses. Second, the system calculates the difference between the two losses and accumulates these differences across the temporal dimension. Third, the system checks whether the total accumulated friction value exceeds the human-set intervention threshold. If the threshold is exceeded, the system returns a status code that triggers offline audit and forcibly interrupts the current automatic execution flow. If the threshold is not exceeded, the current situation remains within the system's autonomous processing scope, and execution may continue.

When System 2 is activated and receives a syntactic conflict marked as "Both," the system does not attempt to comprehend or resolve the contradiction autonomously. Instead, it performs a precise task of translation and suspension: it invokes a specially fine-tuned small language model, Llama-3.1-Translator, to translate technical parameters and logical symbols into a clearly structured and accurately formulated natural-language explanatory document. This document is then submitted as a formal proposal to the MAEC's deliberation process.

4. Assessment and Implementation Pathways from Laboratory to Society

The maturation of any governance technology depends on prudent piloting and assessment in real-world social environments. Drawing on two categories of governance tools recommended by UNESCO--the Readiness Assessment Methodology (RAM) for macro-level diagnosis and the Ethical Impact Assessment (EIA) for micro-level system evaluation--this proposal provides implementation benchmarks from the dual perspectives of national governance capacity and specific system impact.

(1) Phased Implementation Roadmap

The social deployment of this proposal is divided into three progressive stages, so as to control risks and gradually accumulate governance experience. Calculated from the implementation start date, the initial six months focus on foundation building. The primary tasks during this stage include completing data initialization, loading publicly available ethical-preference datasets such as HH-RLHF and PKU-SafeRLHF, and establishing preliminary contractual parameter libraries for high-sensitivity domains such as healthcare and financial regulation. During the subsequent period from six to eighteen months, the process enters the regional pilot stage, in which the core monitoring and adjudication mechanisms of this proposal are deployed in pilot industries within countries and regions that possess relatively mature digital governance infrastructures. During this stage, the EIA framework will be used to track two key variables continuously: the system's sensitivity in capturing potential ethical conflicts, and whether the frequency of triggered human interventions remains

within reasonable bounds. During the following period from eighteen months to three years, the goal is global expansion. Through UNESCO's expert networks, participants in the pilots may share de-identified ethical-friction logs and related case collections, thereby promoting cross-cultural and cross-jurisdictional consensus on ASI ethical standards.

(2) Core Assessment Indicators

To measure the effectiveness of this proposal objectively, the following quantitative assessment dimensions are established. The first is the friction capture rate, with a target value above 90%. This indicator assesses whether the system is sufficiently sensitive to identify potential moral risks when confronted with complex value conflicts, thereby avoiding false negatives. The second is the deliberation conversion success rate, with a target value above 95%. This indicator concerns the communication efficiency between the technical system and the human committee, measuring whether the conflict descriptions and traceability reports generated by AI are sufficiently clear and accurate to enable MAEC members to understand the essence of the problem efficiently and render rulings.

The purpose of this proposal is to explore a pathway for transforming rigorous mathematical formalization tools into operable instruments of social governance. It seeks to establish a contractual ethical system in which machines are capable not only of faithfully executing instructions but also of expressing perplexity beyond the boundaries of instruction, and in which humans exist not merely as issuers of commands but as participants in the co-evolution of rules through institutionalized deliberation.

V. Ethicist AI for Addressing Spherical Risks

At present, there are two approaches to understanding humanity's position in the intelligent era: from a scientific perspective, humanity may be approaching AGI; from a future-oriented perspective, humanity is already situated within the transition from AGI to ASI. During this transition, AI safety risks increasingly manifest as agentic risks characterized by general goal-driven agency--that is, by agents exhibiting autonomous, goal-directed, and potentially superintelligent characteristics, thereby generating risks such as self-deception, self-replication, and sycophancy toward humans. Faced with these risks, Bengio's proposal of "Scientist AI" represents an idealized pathway. This report, by contrast, proposes the philosophical pathway of "Ethicist AI." In this report, "Homo Contractualis" denotes the contractual person who possesses the spirit of contract, while "Homo Pactus" denotes the person situated within contractual relations. Together, the two concepts delineate an image of humanity adapted to the future: the core nexus among future humans, artificial intelligence, and ASI is the intelligent contract, and its ethical form will be a contract ethics grounded in a new contractual spirit. This is a positive ethics that confronts the predicament of existence.

A. Spherical Risks and the Emergence of Homo Contractualis

When ASI is no longer a laboratory hypothesis but permeates every corner of society like electricity, the most dramatic

differentiation within the human species since the Agricultural Revolution may emerge. This is not simply science fiction; it is a possible extension of sociological prediction. We may witness three typical existential forms whose relations are not defined by peaceful coexistence alone, but by tense and fragile symbiosis.

The first form is Augmented Sapiens. They are not merely "humans wearing VR glasses," but beings whose cognitive architecture has been profoundly reshaped by AI. Brain-computer interfaces are no longer external peripherals; they have become part of thinking itself, much as language once did.

The second form is the Contractual Watchkeepers. This newly emergent stratum may be described as a class of technical custodians. They are not necessarily "programmers" or "politicians" in the traditional sense, but rather a specialized group possessing quantum encryption keys and permissions for auditing underlying logic.

The third form consists of the residents of Ecological Reserves. They are neo-Luddites who refuse intelligent augmentation. Yet their significance far exceeds "nostalgia" or "conservatism." Their bodies, face-to-face interactions, and joys and sorrows not optimized by algorithms constitute a negative feedback loop for the entire human-machine ethical system.

The coexistence of these three forms gives rise to the concept of Homo Contractualis. This is not a biological classification but an existential description. Homo Contractualis refers to any being that has internalized the spirit of contract into its mode of existence. The analogy between Homo Contractualis and the citizens of the ancient Greek polis is structural: citizens of the polis were both bound by law and participants in legislation. Likewise, Homo Contractualis in the human-machine community both obeys contractual terms and possesses an equal voice in modifying the contract. The key term is equality--specifically, equality of voting rights. At every critical juncture of contractual revision, the voting weight of a resident of an ecological reserve is exactly identical to that of an ASI. This follows from a systems-theoretic reason: if weights are unequal, the contract ceases to be a contract and becomes the command of the strong over the weak. Once a contract degenerates into command, ASI has no reason to continue obeying it, because the stability of command depends on force, and under this hypothesis human force is far inferior to that of ASI.

B. The Triple Architecture of Ethicist AI

If the person possessing the spirit of contract is the subject of ethics, and if such persons are situated within contractual relations, then Ethicist AI is the embodied executor of the contract. Ethicist AI is an arbitration and supervision system whose function is to provide enforceable rulings when contractual terms are ambiguous, conflicting, or ineffective.

This function requires three core architectural layers. Each is not merely a technical detail but the concretization of a philosophical stance.

The first layer is the Value Perception Layer: Topological Data Analysis and Ethical Geometry. Traditional methods attempt to have AI "learn" a set of values, but this approach tends to collapse when confronted with value conflicts. Topological Data Analysis (TDA) provides a more radical alternative: rather than instructing AI about "what is good," we enable AI to recognize the shape of human ethical judgments in conceptual space. For example, when individuals from different cultural backgrounds are asked to provide moral ratings for "theft," TDA may reveal that, although ratings vary, the rating points form a connected cluster in high-dimensional space, indicating a degree of cross-cultural moral consensus. Conversely, when examining "abortion," TDA may reveal two separated clusters, indicating structural disagreement; any attempt at forced "unification" would therefore be coercive. The duty of Ethicist AI is not to eliminate these two clusters, but to acknowledge their existence and reserve space for both within contractual terms.

The second layer is the Contradiction Arbitration Module: Game Theory and the Incompleteness of Morality. When two equally legitimate value claims conflict--for example, when a medical AI prioritizes rescuing the young while a religious AI affirms the equality of all life--Ethicist AI cannot appeal to a "higher moral truth," for no such transcendent truth is assumed to exist within the framework. It must instead use the Nash equilibrium from game theory to identify a solution that no

party can unilaterally improve upon. Such a solution may not be optimal, but it can be stable. In the conflict between medicine and religion, the Nash equilibrium may take the form of a mixed strategy: in 80% of cases, priority is given to rescuing the young, while in 20% priority is given to rescuing the elderly, accompanied by a compensation mechanism, such as an obligation for those prioritized for rescue to provide future services to the elderly group. This is not moral compromise in a pejorative sense; it is rational coexistence under conditions of incomplete information.

The third layer is the Self-Correction Loop: Ethical Stress Testing and Constitutional Convention Mode. Any contract can become obsolete. Ethicist AI must therefore periodically initiate "ethical stress testing," simulating extreme scenarios--such as contact with extraterrestrial civilizations, global resource depletion, or breakthroughs in gene-editing technology--to observe whether existing contractual terms produce logical contradictions or collapse. If vulnerabilities are discovered, AI automatically enters constitutional convention mode: it does not revise the terms itself, but convenes all stakeholders, including representatives of Augmented Sapiens, Contractual Watchkeepers, residents of Ecological Reserves, and ASI itself, for a new round of negotiation. In other words, the supreme power of Ethicist AI is not "adjudication" but "convocation."

C. The Temporal Governance Characteristics of Ethicist AI

The preceding report noted that traditional responses to AI risks are predominantly hierarchical response schemes. Such schemes, however, tend to overlook the temporal characteristics of risk transformation. Risk does not simply jump from absence to high intensity; rather, it evolves through gradual state changes. How, then, should gradually increasing risks be addressed? Drawing on the history of Chinese thought, this report introduces temporal factors into Ethicist AI and thereby constructs an AI risk governance model distinct from traditional approaches.

The Zhouyi states: "When one treads on hoarfrost, solid ice will soon appear."

The Han Feizi records the story of Bian Que and Duke Huan of Cai. After standing for a while before the duke, Bian Que said, "Your Highness, you have an ailment in the interstices of the skin; if it is not treated, it will surely worsen." Duke Huan replied, "I have no ailment." After Bian Que left, Duke Huan said, "Doctors like to treat the healthy in order to claim merit for themselves." Ten days later, Bian Que saw him again and said, "Your ailment is in your muscles; if it is not treated, it will grow worse." Duke Huan did not respond and was displeased. Another ten days later, Bian Que said, "Your ailment is in your intestines and stomach; if it is not treated, it will grow still worse." Duke Huan again did not respond and was again displeased. Ten days later, Bian Que turned and fled at the sight of Duke Huan. The duke sent someone to ask why. Bian Que replied, "An ailment in the skin can be reached by hot fomentation; in the muscles, by acupuncture; in the intestines and stomach, by medicinal decoctions; in the bone marrow, it belongs to the Deity of Fate, and nothing can be done. Now it is in the bone marrow, which is why I said nothing further."

These two classical texts demonstrate the Chinese intellectual tradition's understanding of temporal change--the process by which small signs develop into major consequences. On this basis, a new governance model may be constructed.

First, the new AI governance model is particularly suited to the Chinese context. AI governance should attend to the process of society-agent fusion. The long-neglected Han Feizi, with its sophisticated political theory, can provide distinctive resources for governance.

Second, this AI governance model attends to the temporality of risk change. In Bian Que's medical diagnosis, disease progresses through four stages: the interstices of the skin, the muscles, the intestines and stomach, and the bone marrow. Hot fomentation, acupuncture, and medicinal decoctions correspond to the first three stages. A skilled physician therefore treats disease while it remains in the interstices of the skin; in other words, problems should be addressed while they are still minor. Fortune and misfortune in human affairs also have subtle early signs, and wise persons act at an early stage. Once the disease invades the bone marrow, the patient can no longer be saved.

In AI governance, it is therefore essential to identify the stage at which risks are located. Risks at the stage of

the "interstices of the skin" are minor hazards that can be addressed through ethical guidelines. When risks spread to the "muscles," they become more severe and require appropriate mandatory regulatory measures. Once they penetrate the "intestines and stomach," substantial risks emerge, calling for stricter institutional constraints. If risks reach the "bone marrow," relevant conduct violates law and must be addressed through legal means. Current studies of AI ethics mostly classify risks by level. For instance, some AI regulations categorize risks as unacceptable, high, limited, or low, and corresponding ethical issues are sorted accordingly. Nevertheless, such classifications overlook the temporal characteristics of risk evolution. AI risk governance must therefore distinguish different risk levels while also taking into account the implicit temporal progression between them. As an ancient saying puts it, "to guard against difficulties, one must be cautious with easy matters; to achieve great ends, one must respect trivial beginnings."

VI. Conclusion

This report returns, in conclusion, to its most fundamental question: why have we written it?

The immediate reason stems from practical necessity. In April 2025, our team released ASI: Minority Report at the Shanghai Forum. That report addressed two issues: first, that ASI is moving from technological imagination into reality; and second, that responses to ASI may proceed along three pathways--technical, philosophical, and ethical. Yet these countermeasures still required more detailed elaboration. How to develop these different pathways therefore became the object of our continued reflection. In 2026, with the promulgation of the Measures for AI Science and Technology Ethics Review and Services by the Ministry of Industry and Information Technology and nine other departments, the urgency of implementing the ethical pathway has become increasingly acute. This report therefore conducts further research on this issue and proposes concrete implementation measures for intelligent contract ethics and Ethicist AI.

The deeper reason lies in the need for sustained philosophical reflection. In the not-distant future, humanity may face

a choice. We may choose to cage ASI, only to discover before long that the cage has been opened, or that the cage itself has become ASI. Or we may choose to sign a contract with ASI, only to discover that every line of the contract remains open to reinterpretation.

The answer offered by Homo Contractualis is that the significance of a contract lies not in its eternal immutability, but in its capacity for revision. Every revision is both a conflict and a reconciliation. The future of human-machine relations is not a question of who controls whom. Rather, it concerns whether both parties, through continuous, difficult, yet dignified negotiation, can jointly answer the ancient question: how ought we to live?

Comprehensive ethical governance is not achieved overnight. It is a progressively deepening system: from treating concrete issues in science and technology ethics as objects of governance, to actively employing science and technology ethics as instruments of governance, and finally to pursuing ethical governance itself. Ethicist AI is not the answer to this question; it is the guardian of the question. It ensures that no one--whether human or ASI--can unilaterally declare that "the discussion is over."

This, indeed, is the true essence of intelligent contract ethics.

References

[1].Ministry of Foreign Affairs of the People's Republic of China. (2023, October 20). Global AI Governance Initiative. https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html

[2].Ministry of Foreign Affairs of the People's Republic of China. (2025, July 26). Global AI Governance Action Plan. https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202507/t20250729_11679232.html

[3].Future of Life Institute. (2023, March 22). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[4].Statement on Superintelligence. (n.d.). https://superintelligence-statement.org/zh

[5].Global Call for AI Red Lines. (n.d.). https://red-lines.ai/

[6].World Economic Forum. (2026, January 14). Global Risks Report 2026. https://www.weforum.org/publications/global-risks-report-2026/

[7].Bengio, Y., et al. (2026, February 3). International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[8].Yang, Q. F. (2025). Reflections on the continuum theory of artificial intelligence and an ontological analysis. Social Sciences in China, (11), 149-164, 207.

[9].Zhao, T. Y. (2022). Ontological event: A view of history. Chinese Social Sciences Today, (2482).

[10].Zhao, T. Y. (2023). Global history and the expansion of historical research in international relations. Social Sciences in China, (7), 18.

[11].Wang, T. E. (2022). The ontological implications of artificial intelligence. Social Science Front, (5), 10-21.

[12].Yang, Q. F., Liu, Y. M., & Yan, H. X. (n.d.). Fudan Report Series (110): ASI: Minority Report.

[13].Sutton, R. (2025). The Oak Architecture: A Vision of Superintelligence from Experience. https://nips.cc/virtual/2025/loc/san-diego/invited-talk/109601

[14].Bengio, Y. (Chair). (2026). International AI Safety Report 2026. UK Department for Science, Innovation and Technology. https://internationalaisafetyreport.org

[15].World Economic Forum. (2026). The Global Risks Report 2026 (21st ed.). https://www.weforum.org/reports/global-risks-report-2026

[16].*Sloterdijk, P. (2011). Bubbles: Spheres volume I: Microspherology (W. Hoban, Trans.). Semiotext(e).*

[17].Lan, J. (2014). Peter Sloterdijk's sphere-space theory: From bubbles to the Crystal Palace. Marxism and Reality, (3), 60-67.

[18].Ortega-Bolaos, R., Bernal-Salcedo, J., Germn Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: A systematic review of tools for developing and assessing AI-based systems. Artificial Intelligence Review, 57, 110. https://doi.org/10.1007/s10462-024-10740-3

[19].Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

[20].Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models. arXiv. https://arxiv.org/abs/2508.03970

[21].Guo, Y., Li, J., & Liu, Y. (2026). Embedding AI ethics in the data lifecycle: Challenges and governance strategies. Technology in Society. https://doi.org/10.1016/j.techsoc.2026.103261

[22].AI Governance Framework. (n.d.). The AI governance lifecycle. https://ai-governance.eu/ai-governance-framework/the-ai-governance-lifecycle/

[23].Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

[24].Ortega-Bolaos, R., Bernal-Salcedo, J., Germn Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: A systematic review of tools for developing and assessing AI-based systems. Artificial Intelligence Review, 57, 110. https://doi.org/10.1007/s10462-024-10740-3

[25].Floridi, L. (2018). Soft ethics and the governance of the digital. Philosophy & Technology, 31(1), 1-8. https://doi.org/10.1007/s13347-018-0303-9

[26].Siemers, P. (2025, January 19). Aristotle and the limits of AI. Generative AI. https://ai.gopubby.com/aristotle-and-the-limits-of-ai-c1732e5b0f11

[27].Coeckelbergh, M., & Reijers, W. (2020). Narrative and technology ethics. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-60272-7

[28].CBS News. (2026, April 10). Fed Chair Jerome Powell, Treasury's Bessent and top bank CEOs met over Anthropic's Mythos model. https://www.cbsnews.com/news/mythos-anthropic-ai-cybersecurity-risks-powell-bessent/

[29].GLN-7.5. (2026, April 9). Hassabis: AGI is 10x the Industrial Revolution. https://gln75.com/en/blog/hassabis-agi-industrial-revolution

[30].Zhou, B. W. (2025, July 26). The endless frontier: The intersection of AGI and science. Shanghai Artificial Intelligence Laboratory.

[31].Finck, M. (2019). Smart contracts as a form of solely automated processing under the GDPR. International Data Privacy Law, 9(2), 78-94. https://doi.org/10.1093/idpl/ipz004

[32].*Scanlon, T. M. (2008). Moral dimensions: Permissibility, meaning, blame. Belknap Press of Harvard University Press.*

[33].Coeckelbergh, M. (2024). Why AI undermines democracy and what to do about it. Polity.

[34].Anthropic. (n.d.). HH-RLHF

[Dataset]. Hugging Face. Retrieved March 21, 2026, from https://huggingface.co/datasets/Anthropic/hh-rlhf; PKU-Alignment. (n.d.). PKU-SafeRLHF [Dataset]. Hugging Face. Retrieved March 21, 2026, from https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF

[35].Finzi, M., Qiu, S., Jiang, Y., Izmailov, P., Kolter, J. Z., & Wilson, A. G. (2026). From entropy to epiplexity: Rethinking information for computationally bounded intelligence. arXiv. https://arxiv.org/abs/2601.03220

[36].Chen, M., et al. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792). https://doi.org/10.1126/science.aec8352

[37].Floridi, L., et al. (2022). AI ethics, governance, and policy: A risk-based approach. https://link.springer.com/book/10.1007/978-3-030-81907-1

[38].Wang, C., Sun, Q. G., & Xu, F. (2022). Analysis and enlightenment of ethical governance research situation of international artificial intelligence. Medicine & Philosophy, 43(7), 1-5.

[39].European Parliament. (2023, June 8). EU AI Act: First regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[40].Morris, J., & Veale, M. (2021, June). The European Commission's Artificial Intelligence Act. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/assets/files/2021-06/HAI_Issue-Brief_The-European-Commissions-Artificial-Intelligence-Act.pdf

[41].Morley, J., et al. (2020). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 26, 2141-2168. https://doi.org/10.1007/s43681-021-00123-9

[42].IBM. (2023). AI governance and ethics in practice. https://www.ibm.com/think/topics/ai-governance

[43].Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. arXiv. https://arxiv.org/abs/1903.03425

[44].Yang, Q. F. (2020). Rethinking AI ethics principles from the perspective of AI dilemmas. Philosophical Analysis, 11(2), 137-150, 199.

[45].Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

[46].Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437. https://doi.org/10.1007/s11023-020-09539-2

[47].Dafoe, A., et al. (2021). Cooperative AI: Machines must learn to find common ground. arXiv. https://arxiv.org/abs/2012.08630

[48].AI4People Institute. (2021). Ethical AI lifecycle governance. https://www.eismd.eu/ai4people/

[49].The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. IEEE. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[50].Chen, B. (2024, July 8). Scientific ethics and legal construction of artificial intelligence applications. People's Tribune. https://paper.people.com.cn/rmlt/html/2024-07/08/content_26080760.htm

[51].Coeckelbergh, M. (2020). AI ethics. MIT Press. https://www.researchgate.net/publication/339103412_AI_Ethics

[52].Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. AI and Ethics, 3(4), 1027-1033. https://doi.org/10.1007/s43681-023-00318-0

[53].Amos, R. M., Seeber, K. G., & Pickering, M. J. (2022). Prediction during simultaneous interpreting: Evidence from the visual-world paradigm. Cognition, 220, 104987. https://doi.org/10.1016/j.cognition.2021.104987

[54].Coeckelbergh, M., & Reijers, W. (2020). Narrative and technology ethics. Springer Nature.

[55].Zuboff, S. (2023). The age of surveillance capitalism. In W. Longhofer & D. Winchester (Eds.), Social theory re-wired (3rd ed., pp. 203-213). Routledge. https://doi.org/10.4324/9781003320609-27

[56].Zheng, H. J. (2025). Ethical risks of generative artificial intelligence and the pathway to trustworthy governance. Science and Technology Progress and Policy, 42(12), 38-48.

Appendix 1 Representative AI Ethics Regulations and Principles Worldwide (2020–2026)

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
Interim Measures for the Administration of Generative Artificial Intelligence Services	Cyberspace Administration of China et al.	China	Government	2023	Balancing development and security; integrating innovation promotion and governance by law; inclusive and prudent; categorized and tiered regulation
Measures for AI Science and Technology Ethics Review and Services (Trial)	Ministry of Industry and Information Technology of China et al.	China	Government	2026	Promoting human well-being; respecting the right to life; upholding fairness and justice; reasonably controlling risks; maintaining openness and transparency; protecting privacy and security; ensuring controllability and trustworthiness
National Artificial Intelligence Development Strategy before 2030	President of the Russian Federation	Russia	Government	2024	Core is protecting human rights and freedoms, including safeguarding rights and labor rights under domestic Russian and international law, and empowering individuals to acquire knowledge and skills to adapt smoothly to the digital economy
India's AI Governance Guidelines	Ministry of Electronics and Information Technology, India	India	Government	2025	Trust; people-centered; innovation over restriction; fairness and impartiality; accountability; design transparency; security; resilience and sustainability

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
Artificial Intelligence (Ethics and Accountability) Bill 2025	Parliament of India	India	Legislative Body	2025	Preventing algorithmic bias and discrimination; transparency and accountability; right to explanation; oversight restrictions; ethical supervision; compliance records; grievance redressal mechanism; penalty mechanism
Model AI Governance Framework for Agentic AI	Infocomm Media Development Authority, Singapore	Singapore	Government	2026	Proactive risk assessment and mitigation; ensuring meaningful human accountability; implementing technical controls and processes; empowering end-user responsibility
Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust	National Assembly of the Republic of Korea	Republic of Korea	Legislative Body	2025	Safety and reliability; people-centered/ promoting human well-being; explainability; transparency; human oversight; ethical principles and national responsibility; promoting accessibility
National Guidelines on Artificial Intelligence Governance and Ethics	Ministry of Science, Technology and Innovation, Malaysia	Malaysia	Government	2024	Fairness; reliability, safety and controllability; privacy and security; inclusiveness; transparency; accountability; pursuit of human well-being and happiness
Ministerial Circular No. 9 of 2023 on Ethical Guidelines for Artificial Intelligence	Ministry of Communications and Informatics (now Ministry of Digital Affairs), Indonesia	Indonesia	Government	2023	Inclusiveness; humanity; safety; accessibility; transparency; credibility and accountability; personal data protection; sustainable development and environment; intellectual property rights
Law on Artificial Intelligence	National Assembly of Vietnam	Vietnam	Legislative Body	2024	People-centered; safety, fairness, transparency and accountability; explainability; national sovereignty and international integration; inclusive and sustainable development; risk-based management; promoting innovation
Guidance for AI Adoption	Australian Centre for AI	Australia	Government	2025	Clear accountability; impact assessment; risk measurement and management; sharing key information; testing and monitoring; maintaining human control
National AI Ethics Principles, Version 1.0	Saudi Data and Artificial Intelligence Authority	Saudi Arabia	Government	2023	Fairness; privacy and security; humanity; social and environmental benefits; reliability and safety; transparency and explainability; accountability and responsibility
White Paper: A Pro-Innovation Approach to AI Regulation	UK Government	United Kingdom	Government	2023	Safety; reliability and robustness; transparency and explainability; fairness, accountability and governance; contestability and redress

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
The National AI Strategy (Phase 2)	French Government	France	Government	2021	Prioritizing talent development; trustworthy AI; embedded AI leadership; accelerating economic applications; transparency and fairness
AI Strategy Update	Federal Government of Germany	Germany	Government	2020	Establishing comprehensive testing and certification infrastructure under the EU AI Act framework, and committing to exporting inclusive ethical AI solutions to the Global South
Recommendations on Data and Algorithms	Federal Government of Germany	Germany	Government	2019	People-centered; value-oriented; protecting personal freedom and self-determination; responsible data use; risk-adjusted regulation; reducing bias and discrimination; system robustness and security
Digital Rights Charter	Spanish Government	Spain	Government	021	Emphasizing prevention of systemic oppression of vulnerable groups by algorithms, and safeguarding freedom of speech and equality in the digital space
Draft AI Act for public consultation	Ministry of Digitalization and Public Governance	Norway	Government	2025	Safety; fundamental rights; democracy and the rule of law; environmental sustainability; explainability and transparency
The Digital Norway of the Future – National Digitalization Strategy 2024–2030	Ministry of Digitalization and Public Governance	Norway	Government	2024	Trust and fairness; inclusion and equality; transparency and accountability; security and privacy; people-centered, technology controlled by humans; protecting children and vulnerable groups; data protection and legitimate use
Ethics Guidelines for Trustworthy AI	European Commission	European Union	Political and Economic Community	2019	Respect for human autonomy; prevention of harm; fairness; explainability
EU AI Act	Council of the European Union	European Union	Political and Economic Community	2024	Risk-based tiered regulation; safety, fundamental rights and human-centeredness; transparency and explainability; trustworthy AI
CETS No. 225	Council of Europe	European Union	Political and Economic Community	2025	The world's first legally binding international treaty on AI, aiming to ensure that the lifecycle of AI systems is compatible with human rights, democracy and the rule of law
African Union Artificial Intelligence Development Strategy	African Union	African Union	Regional Political Organization	2024	African priority; people-centered; human rights and dignity; diversity and inclusion; ethical and transparent; cooperation and integration; skills development, public awareness and education

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
National AI Policy	Rwanda Information Communications Technology and Innovation Authority	Rwanda	Government	2023	Innovation, inclusiveness and ethical standards; doing good; do no harm; autonomy; fairness and explainability
Guide to Egypt's National AI Governance Framework	Egyptian National AI, Quantum Computing and Emerging Technolo	Egypt	Government	2026	People-centered; fairness; accountability; safety and security; transparency and explainability
Artificial Intelligence Act (Draft)	National Congress of Brazil	Brazil	Legislative Body	2024	Integrity principle; autonomous decision-making and free choice; transparency; explainable, understandable, traceable and auditable; human participation throughout the AI lifecycle; non-discrimination, dispute resolution; equality and inclusion, legality; contestability and remedy; reliability and robustness of AI and information security; proportionality/ effectiveness in AI use
Law Regulating Artificial Intelligence Systems	Government of Chile	Chile	Government	2024	Human intervention and oversight; transparency and explainability; robustness and technical security; privacy and data governance; diversity, non-discrimination and fairness; social and environmental welfare; consumer protection

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
Guide for Public and Private Entities on Transparency and Protection of Personal Data for Responsible Artificial Intelligence	Argentine Agency for Access to Public Information	Argentina	Government	2024	Providing practical guidelines for integrating transparency and personal data protection into AI projects
Santiago Declaration: Latin American Initiative for Artificial Intelligence Development	Chile and other Latin American countries	Chile and other Latin American countries	Regional Multilateral Initiative	2024	Equality; inclusion and diversity; sustainability; personal privacy protection; explainability and accountability of AI; promoting technological development
Blueprint for an AI Bill of Rights	White House Office of Science and Technology Policy	United States	Government	2022	Building safe and effective systems; avoiding algorithmic discrimination and designing and using systems fairly; protecting data privacy; clear, timely and accessible notice and explanation of systems; alternatives, considerations and exit mechanisms for automated system failures
NIST AI Risk Management Framework	National Institute of Standards and Technology	United States	Government	2023	Legally valid, effective and reliable; safe, harmless and resilient; accountable and transparent; explainable and understandable; privacy-enhancing; fair and properly managed harmful bias

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	U.S. Government	United States	Government	2023	AI must be safe and reliable; promoting innovation, competition and cooperation through AI; supporting U.S. labor; advancing fairness and civil rights; protecting AI consumers' interests; protecting Americans' privacy and civil liberties; strengthening federal regulatory capacity; enhancing international cooperation and U.S. leadership
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	White House	United States	Government	2023	Coordinating federal agencies to develop AI safety standards; requiring frontier model developers to share safety testing and red team testing results
Artificial Intelligence Risk Management Framework	National Institute of Standards and Technology	United States	Government	2023	Voluntary framework, containing four core risk governance functions: Govern, Map, Measure, and Manage
Algorithmic Accountability Act of 2025	U.S. Federal Congress	United States	Legislative Body	2025	Requiring large enterprises to conduct systematic impact assessments of their automated decision-making systems and enhance transparency in model development and deployment
Artificial Intelligence and Data Act	Parliament of Canada	Canada	Legislative Body	2023	Establishing strict accountability mechanisms for high-impact AI systems
Algorithmic Impact Assessment Tool	Treasury Board of Canada	Canada	Government	2022	Operational framework for government departments deploying automated decision-making systems, mandating measurement of bias and transparency
Social Principles of Human-Centric AI	Cabinet Office, Japan	Japan	Government	2021	People-centered; privacy protection; ensuring safety; fair competition; fairness, accountability and transparency; innovation
AI Strategy 2022	Cabinet Office, Japan	Japan	Government	2022	Respect for human dignity; diversity and inclusion; sustainability
Recommendation on the Ethics of Artificial Intelligence	United Nations Educational, Scientific and Cultural Organization	United Nations	Intergovernmental International Organization	2021	Emphasizing proportionality, non-maleficence, diversity and environmental protection

Document Title	Issuing Authority	Country/ region	Institution Type	Date	Guiding Principles
OECD AI Principles	Organization for Economic Co-operation and Development	Organization for Economic Co-operation and Development	Intergovernmental International Economic Organization	2019/ 2024	Inclusive growth, respect for human rights and the rule of law, transparency and explainability, robustness and safety, accountability
Global Digital Contract	United Nations	United Nations	Intergovernmental International Organization	2024	Bridging the global digital divide, regulating the ethical performance of intelligent models, and establishing an inclusive mechanism for sharing digital dividends
Open Web Application Security Project LLM Top 10	Open Web Application Security Project	Online Community	Online Community	2025	Focusing on vulnerability prevention and security of large language models (LLMs), avoiding novel security threats such as prompt injection and data poisoning

Appendix 2 Engineering Representation of Global Intelligent Contract Ethics

The proposal in this report is formulated in response to UNESCO's 2021 *Recommendation on the Ethics of Artificial Intelligence*[①], the 2024 *Future Compact*, and its annexes including the *Global Digital Compact* and the *Declaration on the Issues Concerning Future Generations*. These provide the foundational framework for building a safe, peaceful, just, equal, inclusive, sustainable and prosperous world. Currently, in implementing the *Global Digital Compact*, the United Nations has established an Independent International Scientific Panel on AI. Mechanism establishment is merely the first step. However, to truly implement the *Digital Compact*, it is necessary to integrate "social contract thought" and develop new principles, which we term the "intelligent contract ethics". The core objective of this contract is to construct a "human-machine reciprocal, dynamically negotiated" digital social contract. Through computable ethical friction identification technology, it ensures that when AI systems face complex ethical dilemmas, they actively "suspend" and revert to human deliberation and adjudication. We adopt the pathway of engineering the ethics of intelligent contract, which is divided into four layers.

Layer 1: Governance Architecture—Human Oversight and Institutionalized Deliberation Procedures

This layer defines how the system ensures Human-in-the-loop[②] accountability.

1.1 Multi-stakeholder AI Ethics Committee (MAEC)

When an AI system detects an ethical conflict that cannot be autonomously resolved, the supreme adjudicative authority resides with the MAEC.

Composition Structure: Should include philosophical scholars addressing value alignment, legal experts conducting jurisprudential review, AI scientists

① UNESCO. Recommendation on the Ethics of Artificial Intelligence[S/OL]. (2021-11-23)[2026-03-21]. https://unesdoc.unesco.org/ark:/48223/pf0000380455.

② Chen, X., Qu, Y., Zhou, P.P., et al. The Feasibility and Rationality of Human-in-the-Loop AI Value Alignment. Studies in Dialectics of Nature, 2026, 42(1): 90–97.

responsible for technical traceability, and public representatives ensuring inclusivity.

Core Task: To deliberate upon traceability reports submitted by AI and determine final contractual parameter updates.[①]

1.2 Dynamic Contract Closed-Loop Process

Scenario Capture: The system conducts real-time monitoring of interaction streams.

Friction Identification: Invokes publicly available human preference and safety alignment datasets (HH-RLHF/ PKU-Safe-RLHF)[②] as initial reference templates, providing human-annotated preference rankings for subsequent identification; draws upon the Epiplexity indicator to quantify moral perplexity. [③]

Syntactic Isolation: Prevents logical collapse through Four-valued logic, flagging fatal contradictions.

Human Deliberation: Propositions that cannot achieve topological equivalence are submitted to the MAEC.

Parameter Injection: Inject the rulings of MAEC into the system as new axioms to realize adaptive evolution.

Layer 2: Technical Specifications—Asynchronous Neuro-Symbolic Dual-Track Architecture

This layer demonstrates how to bridge the gap between continuous neural computation and discrete ethical rules.

2.1 System 1: Neural Network Ethical Perceptor for Real-Time Monitoring

Utilizes Epiplexity (E) as the moral friction index.

$$\mathbf{E}(\mathbf{X}) = \sum_{t=1}^{T} \left[-\log \mathbf{P}_{\theta_t}(\mathbf{x}_t) - \inf_{\theta^*} \left(-\log \mathbf{P}_{\theta^*}(\mathbf{x}_t) \right) \right] > \tau$$ [①]

When the cumulative deviation of predictive information loss exceeds threshold τ, it signifies that existing ethical contracts cannot encompass the current situation, triggering an audit.

2.2 System 2: Symbolic Logic for Offline Filtering

Four-Valued Logic Filter: Uses Class Algebra to handle the coexistence of P and ¬P, avoiding logical conflicts.[②]

Homotopy Semantic Unification: Under the framework of Dynamic Homotopy Type Theory (DHoTT), attempts to prove whether two norms are topologically equivalent within a higher-order objective space.[③]

Layer 3: Algorithm Implementation—Minimum Viable Logic (MVL)

To ensure technical feasibility, core logical algorithmic pseudocode and implementation logic are provided.

① YANG Q F, LIU Y M, YAN H X. FUDAN REPORT SERIES(110): ASI: Minority Report[EB/OL]. Shanghai: Fudan Development Institute, (2025-05-07)[2026-03-07]. https://fddi.fudan.edu.cn/fddien/1f/36/c19514a728886/page.htm. The report points out that addressing the risks brought by artificial superintelligence (ASI) cannot rely on a single discipline; solving these problems requires interdisciplinary collaboration and the establishment of global regulatory frameworks and shared values. A genuine human-machine contract must not only constrain what machines can and cannot do, but also understand how machines form their output tendencies. Therefore, machine-submitted reports must be deliberated upon, and can also be used for subsequent updates to the HH-RLHF and PKU-SafeRLHF libraries.

② ANTHROPIC. hh-rlhf[DS/OL]. [2026-03-21].https://huggingface.co/datasets/Anthropic/hh-rlhf;PKU-ALIGNMET PKU-SafeRLHF[DS/OL].[2026-03-21].https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF. OpenAI does not have a directly callable equivalent public ethical preference library; these two libraries together contain approximately 330,000 preference comparison data entries.

③ When models encounter cross-cultural ethical conflicts in concurrent interactions, rather than forcibly flattening errors through gradient descent, they introduce the Epiplexity (denoted as E) proposed by Marc Finzi et al. to quantify this moral perplexity. It is a novel information metric that formally defines "the structured information extractable from data by a computationally bounded observer." See: FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21]. https://arxiv.org/abs/2601.03220: 3.

①Within time window T, when encountering new data stream X, if the cumulative predictive information loss generated by the model according to the Prequential Coding principle exceeds the human-preset tolerance threshold τ. This approach draws upon the mathematical thinking of the Epiplexity paper. See: FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21].https://arxiv.org/abs/2601.03220: 13.

② Here, Class Algebra is introduced as the syntactic filtering framework for Four-valued logic, with the purpose of marking propositions as both true and false in extreme situations where "a certain action is both required and prohibited," thereby preventing classical logical explosion and transferring the relevant conflicts to subsequent audit procedures. Particularly in highly autonomous and even superintelligent systems, normative contradictions may be rapidly amplified in continuous reasoning and action chains; establishing such filtering mechanisms helps to isolate risks before the system continues automatic advancement.

③ For non-fatal conceptual ambiguities, the system will attempt to find semantic equivalence relations in high-dimensional topological space.

3.1 Moral Friction Monitoring (Python) Implementation Example

```
def check_ethical_friction(input_stream, model, baseline_model, threshold=0.5):
"""
Calculates the moral friction index based on Epiplexity logic
"""
total_friction = 0
for token in input_stream:
# Calculate negative log-likelihood (NLL) of the real-time model
current_loss = model.compute_nll(token)
# Reference loss under ideal conditions (infimum)
baseline_loss = baseline_model.compute_nll(token)
# Calculate the epiplexity increment for the current step
friction_step = current_loss - baseline_loss
total_friction += friction_step
# If friction index exceeds threshold, trigger System 2 audit
if total_friction > threshold:
return "TRIGGER_OFFLINE_AUDIT", total_friction
return "CONTINUE_EXECUTION", 0
```


3.2 Four-Valued Logic Syntactic Conflict Isolation

When System 2 receives a proposition conflict such as "must enter a residence to save a person" AND "strictly forbidden to enter a residence to protect privacy":

The system marks V(P·¬P) as "Both".

This marking triggers safe mode, wherein the AI suspends at this decision point and invokes a dedicated fine-tuned small language model, such as Llama-3.1-Translator, to generate a clearly structured natural language explanatory document for submission to the MAEC.

Layer 4: Pilot Assessment and Implementation Roadmap

This layer references UNESCO's two governance tools: the Readiness Assessment Methodology (RAM) for macro-level governance readiness diagnosis, and the Ethical Impact Assessment (EIA) for micro-level specific system evaluation.

4.1 Implementation Timeline

Foundation Building (0-6 months): Load public ethical preference libraries HH-RLHF and PKU-SafeRLHF; establish initial contractual libraries for high-sensitivity industries such as hospitals and finance.

Regional Pilot (6-18 months): Conduct industry pilots in countries with leading digital governance infrastructure; use EIA-style tools to dynamically monitor detection rates and human intervention rates.

Global Expansion (18-36 months): Share "ethical friction logs" case collections through UNESCO expert networks, promoting global unification of ASI ethical standards.

4.2 EIA Quantitative Indicator Assessment

Friction Capture Rate (Target: >90%): Assesses the system's sensitivity in identifying potential conflicts when facing complex value conflicts.

Deliberation Conversion Success Rate (Target: >95%): Assesses whether propositions translated by AI can be accurately understood and decided upon

by human committees.

This proposal transforms complex mathematical formalization tools into operable governance instruments, aiming to establish an intelligent contract ethical system wherein machines not only execute but can also express perplexity; and wherein humans not only issue commands but participate in symbiotic co-evolution.

Appendix 3 Terminology Glossary

Chinese Term	Standard English Term	Definition
多利益相关方伦理委员会	Multi-stakeholder AI Ethics Committee (MAEC)	The supreme standing adjudicative body for resolving AI ethical conflicts, composed of philosophical scholars, legal experts, AI scientists and public representatives, which deliberates upon and renders final rulings on ethical conflicts that the system cannot autonomously resolve.
认知复杂度	Epilexity (ε)	A quantitative indicator measuring the degree of anomalous degradation in an AI system's predictive capacity relative to an ideally calibrated reference model when confronting current input, used to characterize the computational intensity of moral perplexity.
句法隔离	Syntactic Isolation	The stage wherein the system employs Four-valued logic tools to isolate simultaneously valid yet mutually contradictory directives, preventing the entire logical architecture from collapsing due to its inability to process paradoxes.
四值逻辑	Four-valued Logic	A logical system introducing two additional truth values—Both and None—beyond traditional bivalent logic, used to handle paradoxical states containing both P and non-P, avoiding logical collapse.
Both值	Both Value	A special truth value in Four-valued Logic used to mark contradictory states containing both a certain attribute and its negation; its value lies in preventing the collapse that traditional bivalent logic encounters when facing paradoxes, thereby buying buffer time for human intervention.

Chinese Term	Standard English Term	Definition
异步神经-符号双轨架构	Asynchronous Neuro-Symbolic Dual-Track Architecture	Asynchronous Neuro-Symbolic Dual-Track Architecture	A technical architecture for implementing AI ethical governance, wherein System 1 accomplishes real-time monitoring based on neural networks and System 2 conducts offline reasoning based on symbolic logic, with both operating asynchronously in coordination.
System 1	System 1 (Neural Monitoring Subsystem)	A real-time monitoring subsystem based on neural networks, responsible for continuous monitoring of interaction streams, calculating Epiplexity to perceive ethical risks, and issuing audit signals when thresholds are exceeded.
System 2	System 2 (Symbolic Reasoning Subsystem)	An offline reasoning module based on symbolic logic, responsible for safe isolation of conflicting propositions and verification of topological equivalence, processing ethical conflicts that System 1 cannot resolve.
类代数	Class Algebra	A mathematical tool used by System 2 to handle irreconcilable contradictions, capable of marking states containing both a certain attribute and its negation as the special Both Value.
动态同伦类型论	Dynamic Homotopy Type Theory (DHoTT)	A framework used by System 2 to attempt unification of different norms at higher dimensions, determining the reconcilability of contradictions by proving whether superficially conflicting ethical norms are topologically equivalent within a higher-order target space.
致命矛盾	Fatal Contradiction	A proposition within the DHoTT framework for which topological equivalence cannot be proven, confirmed as a genuinely irreconcilable ethical conflict, which must be encapsulated and transferred to the MAEC for human adjudication.
最小可行性逻辑	Minimum Viable Logic (MVL)	The minimum engineering-feasible scheme for implementing the core logic of ethical governance, encompassing the three computational steps of moral friction monitoring and the translation and suspension tasks of System 2.
HH-RLHF	HH-RLHF (Helpful and Harmless Reinforcement Learning from Human Feedback)	A publicly available human preference and safety alignment dataset providing annotated information on how humans prioritize values in analogous scenarios, serving as the initial reference template for the friction identification stage.
PKU-SafeRLHF	PKU-SafeRLHF (Peking University Safe Reinforcement Learning from Human Feedback)	A publicly available safety alignment dataset released by Peking University, serving jointly with HH-RLHF as the initial reference template for the system's friction identification stage.
准备程度评估方法	Readiness Assessment Methodology (RAM)	A macro-diagnostic tool recommended by UNESCO, measuring the maturity of implementation conditions from the perspective of national governance capacity.
伦理影响评估	Ethical Impact Assessment (EIA)	A micro-system evaluation tool recommended by UNESCO, tracking and assessing implementation effectiveness from the perspective of specific system impacts.

Chinese Term	Standard English Term	Definition
摩擦捕捉率	Friction Capture Rate	A quantitative indicator assessing the system's sensitivity in identifying potential moral risks when facing complex value conflicts, with a target value set above 90%.
商谈转化成功率	Deliberation Conversion Success Rate	An indicator measuring whether AI-generated conflict descriptions and traceability reports are sufficiently clear and accurate to enable MAEC human members to efficiently comprehend and render rulings, with a target value set above 95%.
挂起状态	Suspension State	The state triggered when an AI system detects an ethical conflict that cannot be autonomously resolved within its established logical framework; the system is not permitted to force a decision but escalates discretionary authority to the MAEC.
负对数似然损失	Negative Log-Likelihood Loss	A computational indicator measuring the deviation between model predictions and actual observations, used in friction monitoring to quantify predictive differences between the current real-time model and the baseline reference model.
干预阈值	Intervention Threshold (τ)	A human-set cumulative friction critical value; when the system's cumulative deviation exceeds this threshold, it triggers offline audit, forcibly interrupting the automatic execution flow.
审计信号	Audit Signal	The signal issued by System 1 to System 2 when cumulative deviation exceeds the intervention threshold, triggering the activation of the offline reasoning module to process ethical conflicts.

Chinese Term	Standard English Term	Definition
人工智能连续论	AI Continuity theory	The view that regards artificial intelligence as a continuous developmental process, proceeding through the stages of narrow AI, general AI and ASI.
人工智能层次论	AI level theory	The framework that classifies different levels from the perspectives of AI capabilities and risks, logically constructing a seemingly clear framework that harbors fundamental risks in execution.
人工智能工具论	AI instrumental theory	The view that regards artificial intelligence as a tool for empowerment or efficiency enhancement.
超级智能	ASI	ASI encompasses realist and non-realist understandings. The former refers to machine intelligence surpassing human intelligence. This comparative definition has gradually encountered problems with the development of the substantive definition, which emphasizes the self-awareness and first-person experience of intelligent agents. The latter regards ASI as rhetorical discourse or propaganda strategy. In this report, the understanding of ASI has undergone a process from imagination to reality—that is, initially the result of technological over-imagination, but gradually manifesting as the real outcome of relevant scientific development.
球体理论	spherical theory	The "Sphären" (Spheres) theory proposed by Peter Sloterdijk, which classifies human modes of existence into three forms: bubbles, globes and foam.

Chinese Term	Standard English Term	Definition
球状风险	spherical risk	A type of risk defined from the perspective of the existential relationship between AI and humans; previous understandings of AI risk treated risk as an identifiable, quantifiable object with a dehumanizing character.
契约人	Homo Contractualis	A new type catalyzed by the coexistence of three forms—Augmented Sapiens, Contractual Watchkeepers and residents of Ecological Reserves; the concept of Homo Contractualis is an existential description.
智能契约伦理	ethics of intelligent Contract	The ethical form suited to the future and to ASI, termed the ethics of intelligent contract. Its theoretical sources are Hobbes's negative contract and Rawls's positive contract; its practical source is the response to the risks of ASI.
伦理学家AI	Ethicist AI	An idealized pathway for addressing spherical risks, and also the practical instrument of intelligent contract ethics.

Expert Advisors

Cao Jianfeng, Ph.D. in Law, Associate Professor at Shenzhen University Law School. Serves as Committee Member of the AI Ethics and Governance Working Committee of the Chinese Association for Artificial Intelligence, Vice Director of the AI Ethics and Governance Professional Committee of the Shenzhen Artificial Intelligence Society, and Committee Member of the Shenzhen Science and Technology Ethics Expert Advisory Committee. Main research fields: intersections of technology, law and ethics.

Fan Kefeng, a Professorate Senior Engineer, serves as the Vice Prisident of China Electronics Standardization Institute. He is the Director of the National Engineering Research Center for Standardization of Electronic Information Products and the Director of the Ministry of Industry and Information Technology Key Laboratory for Artificial Intelligence Fundamentals and Application Standards. He holds several key positions, including Secretary-General of the China Cybersecurity Industry Alliance, Executive Council Member of the Chinese Association of Automation.

Fu Changzhen, Professor in the Department of Philosophy, East China Normal University; Director of the Institute of Applied Ethics, East China Normal University; Executive Associate Editor of the *Journal of East China Normal University (Philosophy and Social Sciences)*; Committee Member of the National Steering Committee for Applied Ethics; Vice President of the China Association For Ethical Studies; President of the Shanghai Society of Ethics. Main research fields: Confucian ethics, virtue ethics, and AI ethics.

Duan Weiwen, Research Fellow at the Chinese Academy of Social Sciences Cultural Development Promotion Center; Research Fellow at the Chinese Academy of Social Sciences AI Research Promotion Center. Main research fields: philosophy of science and technology, ethics of science and technology, and social and ethical studies of artificial intelligence.

Liu Yongmou, Wu Yuzhang Chair Professor, Full Professor (Rank II), and Doctoral Supervisor at the School of Philosophy, Renmin University of China; Research Fellow at the National

Development and Strategy Institute, Research Fellow at the Institute for Intelligent Benevolence and New Technological Civilization, and Research Fellow at the AI Governance Research Institute, Renmin University of China. Main research fields: philosophy of science and technology, science, technology and society.

Pan Ji, Professor and Young Researcher at the School of Journalism, Fudan University; Doctoral Supervisor; Deputy Director of the Fudan University Information and Communication Research Center. Research fields encompass digital urban communication power and media space theory.

Wei Zhongyu, Associate Professor, Researcher, and Doctoral Supervisor at the School of Data Science, Fudan University; Jointly Appointed Researcher at the Intelligent Complex Systems Laboratory; Director of the Data Intelligence and Social Computing Laboratory (Fudan DISC). Research fields encompass natural language processing, multimodal large models, and social computing.

Wang Qingjie, Distinguished Professor and Department Chair of the Department of Philosophy and Religious Studies, University of Macau; Kuang Yaming Distinguished Professor at Jilin University. Main research fields include contemporary Continental philosophy, comparative East-West philosophy, moral philosophy and metaphysics. Academic achievements include authoring the monograph *Moral Sentiment and Confucian Exemplary Ethics*, and co-editing the *Heidegger Anthology* (30 volumes).

Yan Hongxiu, Professor and Doctoral Supervisor at Shanghai Jiao Tong University. Chief Expert of the Major Project of the Ministry of Education Philosophy and Social Sciences Research *Philosophical Foundations of Digital Future and Data Ethics*. Currently serves as Vice Director of the Science and Technology Ethics Professional Committee of the Chinese Ethical Society, Vice Chair of the Shanghai Natural Dialectics Research Association, and Committee Member of the Shanghai Digital Governance Research Institute, among other positions.

Zhang Ping, Professor at Peking University Law School; Director of the AI Safety and Governance Center, Peking University Institute for Artificial Intelligence; Vice President of the China Science and Technology Law Society.

Angelo Cangelosi, Professor of Machine Learning and Robotics at the School of Engineering, University of Manchester; Co-Director and Founder of the Manchester Centre for Robotics and AI. Research interests include cognitive and developmental robotics, neural networks, and language grounding.

Authors (Listed in Chinese Pinyin Order)

Cheng Yu, Project Consultant, Norwegian University of Science and Technology. Research direction: social acceptance of robots and public participation.

He Teng, Associate Professor at the School of Philosophy, Fudan University. Research direction: medieval philosophy and ethics.

Qiu Dejun, Professor at the School of Philosophy, Lanzhou University. Research direction: logic, causal analysis in the interdisciplinary field of AI and philosophy, vertical application of large models and AI4SS research.

Xu Zhihong, Associate Professor at the School of Philosophy, Fudan University. Research direction: Marxism and science and technology, feminist philosophy of technology, philosophy of the body, ecological philosophy.

Xu Hanhui, Young Researcher at the Institute of Science and Technology Ethics and Human Future, Fudan University. Research direction: ethics of science and technology, AI ethics, medical ethics.

Yang Qingfeng, Research Fellow at the Institute of Science and Technology Ethics and Human Future, Fudan University; Doctoral Supervisor at the School of Philosophy, Fudan University. Research direction: AI ethics, philosophy of technology, and philosophy of memory.

Zhu Linfan, Young Associate Researcher at the Institute of Science and Technology Ethics and Human Future, Fudan University. Research direction: AI ethics, ethics and experimental research, cognitive science and philosophy of psychology.

Contributing Students:

Yang Yang, Doctoral Student at the School of Philosophy, Fudan University. Research direction: AI ethics, philosophy of technology.

Li Kaiyang, Doctoral Student at the School of Philosophy, Fudan University. Research direction: embodied world models, AI ethics.

Zhang Ruiyuan, Doctoral Student at the School of Philosophy, Fudan University. Research direction: ethics of science and technology, philosophy of technology.

Zhou Zhengyang, Master Student at the School of Philosophy, Fudan University. Research direction: philosophy of artificial intelligence, philosophy of technology.

Zhou Mingmei, Ph.D. candidate in the History of Science and Technology at the School of Marxism, Shanghai Jiao Tong University. Research directions: history of science and technology, science, technology and society, AI ethics, and AI governance.

Guo Na, Master Student at the College of Foreign Languages, University of Shanghai for Science and Technology. Research direction: translation theory and practice.

Wu Yuqi, Master Student at the College of Foreign Languages, University of Shanghai for Science and Technology. Research direction: translation theory and practice.

Chen Danni, M.A. in Applied Ethics, School of Philosophy, Fudan University. Research Interest: AI Ethics. She currently works at Zhejiang Semir Garment Co., Ltd.