German Law Journal (2025), **26**, pp. 1117–1148
doi:10.1017/glj.2026.10185

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

From Principle to Practice: Value Alignment in AI Ethics and Governance

Jianfeng Cao

Law School, Shenzhen University, China
jeff.cao.jfc@gmail.com

(Received 30 May 2025; accepted 10 July 2025)

Abstract

As China rapidly advances in AI innovation and development, especially in frontier AI, its regulatory and ethical frameworks are under increasing pressure to ensure that technological progress aligns with human interests and societal values. This Article argues that AI value alignment—the process of ensuring AI systems act in accordance with human values, norms, and ethical principles—should be adopted as a strategic pillar in China's evolving AI governance architecture. While China has already established a comprehensive legal, ethical, and self-regulatory landscape to address AI risks, these mechanisms often rely on reactive enforcement and external compliance. In contrast, AI value alignment offers a proactive, intrinsic approach that embeds safety and ethical constraints directly into AI systems, making them safer, more trustworthy, and responsive to human needs.

This study begins by mapping China's current AI governance landscape, including national legislation such as the Cybersecurity Law, Personal Information Protection Law, and a growing set of regulations targeted at algorithms and generative AI. It also evaluates China's normative commitments, such as the "human-centric" and "tech for good" principles articulated in national policy documents, and the increasing role of corporate self-regulation among major technology firms. While commendable in scope and ambition, these governance mechanisms often fall short in ensuring that AI behavior aligns with safety constraints and ethical intent—particularly when AI systems (such as agentic AI) become more autonomous and capable. This gap highlights the urgent need for a systematic value alignment strategy.

The Article then delves into the conceptual and technical foundations of AI value alignment, identifying both engineering challenges—such as reward misspecification, data bias, and model deception—and normative dilemmas, including moral pluralism, value aggregation, and dynamic ethics. Special attention is paid to frontier models like large language models and artificial general intelligence (AGI), which pose alignment challenges at a scale previously unseen. Drawing on contemporary alignment techniques such as RLHF (Reinforcement Learning from Human Feedback) and principle-based alignment, such as Anthropic's Constitutional AI, the Article explores their limitations and calls for a more diversified, interdisciplinary, and forward-looking alignment research agenda.

Finally, the Article offers a roadmap for operationalizing AI value alignment across three key governance domains: Law and regulation, ethical norms, and industry self-regulation. Recommendations include the incorporation of alignment assessments into regulatory filings, the development of technical standards for value alignment and ethics-by-design guidelines, and institutional investments in safety and alignment research. The Article concludes by asserting that value alignment is not merely a technical safeguard but a governance imperative for the age of autonomous AI and agentic AI. By integrating alignment into its AI governance strategy, China can not only enhance domestic safety and public trust but also better coordinate with global AI ethics and safety initiatives—ultimately contributing to the shared goal of human-aligned and beneficial artificial intelligence.

Keywords: Frontier AI; AI Governance; AI Safety; Value Alignment; Ethics by Design; Responsible AI

A. Introduction

The twenty-first century will be defined by artificial intelligence (AI). In other words, AI will profoundly alter the world we know, as well as what we see, think, and do. Since 1950, when Alan Turing—hailed as the father of AI—raised the fundamental question of "Can machines think?,"[1] researchers have been exploring how to build intelligent machines capable of "thinking" and "acting" like humans. After more than seventy years of exploration and development, particularly in the last decade, thanks to a convergence of big data, computational power—cloud computing, AI-specific chips, and so forth—, machine learning algorithms—deep learning, reinforcement learning, and others—, open-source software frameworks, and other technical elements—the AI field has made unprecedented progress. Moreover, there are no signs of it stopping or slowing down. Google's chief scientist Jeff Dean has referred to the decade since 2010 as the "golden decade" of deep learning.[2] It could be said that advances in and applications of AI technologies not only enable various products to become increasingly autonomous[3] but also propel society from the internet era into an AI society or "algorithmic society,"[4] where algorithms, robots, and AI agents play a central role in economic and social decision-making.[5]

Although over the past decade and more, AI systems driven by deep learning have made one groundbreaking breakthrough after another—often outperforming humans in many tasks—their capabilities have largely been confined to specific domains, with very limited generality. This limitation has been particularly evident in the realm of human language, where progress has long lagged behind other areas in which humans excel. That remained true until late 2022, when ChatGPT, an AI chatbot application powered by a large language model (LLM), emerged, introducing a new developmental paradigm to the AI field: The swift rise of generative artificial intelligence (Generative AI). Generative AI models can learn not only human language but also computer code, images, audio, video, molecules, and many other data types. Models such as GPT-4, Midjourney, Qwen, and DeepSeek V3, both in China and abroad, can autonomously generate text, code, images, audio, video, and more based on user-provided prompts. Their capabilities are now close to, or even surpass, human levels.

These generative AI systems, which may feature hundreds of billions, trillions, or even larger numbers of parameters, are also referred to as foundation models[6] or large AI models. As deep learning models trained on enormous, diverse datasets, they are showing, for the first time, powerful, wide-ranging, and general capabilities in language understanding and content generation—heralding a major turning point in the AI field, namely the shift from specialized AI to general AI. For example, ChatGPT stands out due to its versatility and extensive skill set: By self-learning from massive amounts of data, it has captured knowledge far beyond that of any individual, enabling it to handle an array of tasks ranging from writing stories and summarizing texts, to answering professional questions, translating languages, coding, conducting logical reasoning, explaining concepts, making decisions, offering customer support, and using other tools or services. GPT-4, the large model behind ChatGPT, is even viewed by some as an early version of

[1]Alan M. Turing, *Computing Machinery and Intelligence*, 49 MINDS 433, 433–35 (1950), https://courses.cs.umbc.edu/471/papers/turing.pdf.

[2]Jeffrey Dean, *A Golden Decade of Deep Learning: Computing Systems & Applications*, 151 DAEDALUS 58, 59–62 (2022).

[3]*See* Pangambam S., *How AI Can Bring on a Second Industrial Revolution: Kevin Kelly (Transcript)*, THE SINGJU POST (Aug. 13, 2019, 3:00 AM), https://singjupost.com/how-ai-can-bring-on-a-second-industrial-revolution-kevin-kelly-transcript/?singlepage=1.

[4]Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1217–19 (2017).

[5]*See* Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 89–90 (2014).

[6]*See* Elliot Jones, *What Is a Foundation Model?*, ADA LOVELACE INST. (July 17, 2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

Artificial General Intelligence (AGI),[7] showing signs of intelligence comparable to that of humans. Furthermore, the most critical feature of large models is their emergent abilities—that is, once a model's scale—in parameters and weights—becomes large enough, new capabilities arise that smaller models do not possess, such as arithmetic skills.[8] According to this logic, merely making AI models bigger can endow them with the ability to do a variety of tasks, becoming more autonomous, practical, flexible, and general-purpose tools. Thus, the development of ChatGPT has not only sparked a new wave of AI in the commercial sphere but also rekindled aspirations for achieving AGI at a level on par with human intelligence. In short, there are high expectations for Generative AI: It is viewed as the most transformative AI technology in over a decade, ushering in a new "golden era"[9] of AI that will fundamentally change everything from commerce and science to society itself.[10]

Despite the fact that frontier AI models, exemplified by large language models, continue to advance AI technology at the cutting edge—bringing us closer to AGI or even superintelligence—the rapid development of Generative AI has triggered debates such as "AI accelerationism vs. AI value alignment."[11] It has also led to a series of legal, ethical, and safety concerns, including those related to personal information protection and privacy violations, algorithmic discrimination, misinformation and deepfakes, intellectual property rights and ownership, safety risks, misuse of technology, accountability, technological unemployment, human-AI relations, autonomous weaponization, and the risk of technology spiraling out of control. Moreover, whether more powerful, more general AI models—so-called "frontier AI"—could pose catastrophic or existential risks has become a growing concern. Many experts think if future AI systems surpass human intelligence, they might pose unprecedented challenges for humanity's ability to maintain control over such technologies. Notable experts and scholars in the AI field have been predicting the timeline for AGI and imagining the transformations and impacts it may bring in the next decade or more. For example, Geoffrey Hinton, widely regarded as the "godfather of AI," has issued stark warnings about the existential risks posed by artificial intelligence. He estimates a 10% to 20% chance that AI could lead to human extinction within the next thirty years, a figure he revised upward from earlier predictions.[12] The profound and revolutionary changes and impacts

[7] *See* Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro & Yi Zhang, Sparks of Artificial General Intelligence: Early Experiments with GPT-4, at 92 (2023), https://arxiv.org/pdf/2303.12712.

[8] *See* Boaz Barak, *Emergent Abilities and Grokking: Fundamental, Mirage, or Both?*, Windows On Theory (Dec. 22, 2023), https://windowsontheory.org/2023/12/22/emergent-abilities-and-grokking-fundamental-mirage-or-both/.

[9] Davos-Klosters, *Satya Nadella Says AI Golden Age Is Here and 'It's Good for Humanity'*, World Econ. F. (Jan. 18, 2023), https://www.weforum.org/press/2023/01/satya-nadella-says-ai-golden-age-is-here-and-it-s-good-for-humanity/.

[10] *See* Paul Daugherty, Bhaskar Ghosh, Karthik Narain, Lan Guan & Jim Wilson, A New Era of Generative AI for Everyone: The Technology Underpinning ChatGPT Will Transform Work and Reinvent Business 3 (Accenture 2023), https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf.

[11] AI accelerationism is the idea or ideology advocating for the rapid and unrestrained development of artificial intelligence technology. Proponents argue that quickly advancing AI can lead to major breakthroughs and substantial societal benefits, such as increased productivity, economic prosperity, faster technological progress, and accelerated scientific discovery. AI value alignment is the idea or principle of designing and deploying AI systems that align closely with human values, ethics, norms, and preferences. This concept emphasizes ensuring that advanced AI systems do not act against human interests or well-being, intentionally or unintentionally. Advocates of value alignment seek to embed moral and ethical considerations into AI development and build ethical AI systems, aiming to mitigate risks such as bias, discrimination, manipulation, or even catastrophic misalignment_in which an AI system pursues harmful goals due to poorly specified objectives. While supporting technological progress, value alignment stresses safety, responsibility, and thoughtful governance in AI development. *See* De Kai, *Should A.I. Accelerate? Decelerate? The Answer Is Both.*, The N.Y. Times (Dec. 10, 2023), https://www.nytimes.com/2023/12/10/opinion/openai-silicon-valley-superalignment.html?hpgrp=c-abar&smid=li-share.

[12] Dan Milmo, *'Godfather of AI' Shortens Odds of the Technology Wiping Out Humanity Over Next 30 Years*, The Guardian (Dec. 27, 2024), https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years.

of this AI revolution may be impossible to fully anticipate—or may be seriously underestimated. However, at the very least, we should not underestimate the long-term influence of AI technology.

In the author's view, AI and related technologies are introducing unprecedented risks and challenges for individuals and society that go beyond those posed by any previous technological developments. These risks can be categorized into three main areas: Delegation of decision-making to AI systems, the use of AI for emotional surrogacy, and the potential for AI to be used for human enhancement.[13]

First, as AI and robots become increasingly sophisticated and autonomous, they are being used more and more to assist or even replace humans in making decisions and taking actions across a wide range of economic and social domains. The most prominent example would be AI agents, which can take actions on its own. While this delegation of decision-making to autonomous AI systems can bring benefits in terms of accuracy, efficiency and scale, it also carries significant new risks. These include the potential for AI systems to "hallucinate" or generate false information, to perpetuate societal biases through flawed algorithms and datasets, to optimize for goals that are misaligned with human values, to displace human workers and increase technological unemployment, and to pose existential threats to humanity if AI systems were to advance beyond our control. Given the high stakes involved, we may need to seriously consider whether there are certain types of decisions and human affairs that should never be outsourced to AI, no matter how capable the systems become.

Second, AI and robotic systems are beginning to penetrate deeply into the emotional realm of human life by providing forms of emotional companionship and support. While these "emotional surrogates" may offer certain benefits, such as combating loneliness and supporting elderly people, they also risk diminishing genuine human connections and shifting social norms in problematic ways. As AI becomes more emotionally intelligent and as people grow increasingly comfortable forming relationships with AI agents and AI companions, it could have major detrimental effects on human social interaction and solidarity.[14] Therefore, it is critical that we consider how to define the ethical boundaries of these new human-machine relationships. The key principle, in the author's view, is that human-AI interactions must ultimately serve to reinforce and enrich authentic human bonds, as such human connections will only become more precious in an age where AI is ubiquitous.

Finally, as AI systems become more integrated with human bodies and brains through technologies like brain-computer interfaces (BCI), it could usher in a "post-human era" of human development where the boundaries between humans and machines start to blur.[15] The use of AI for various forms of cognitive and physical enhancement may offer powerful ways to augment human abilities, but also risks exacerbating social inequalities as enhancements may only be accessible to certain segments of society. This also raises profound philosophical questions about the future of human identity and agency in a world where human capabilities are increasingly shaped by AI and related systems. Will there still be a meaningful distinction between human and machine intelligence, or will they merge together? What will it mean to be human in such an age? These are questions we must grapple with now, before such a future unfolds.

In conclusion, the author argues that the delegation of decision-making, emotional surrogacy, and human enhancement applications of AI are game-changing capabilities that raise ethical quandaries beyond the scope of any previous technologies. AI's ability to autonomously make high-stakes decisions, to form emotionally significant relationships with humans, and to literally reshape

[13]Jianfeng Cao, *Human-AI Alignment in the Era of Large AI Models*, Chinese Soc. Scis. Today (Oct. 29, 2024), https://www.cssn.cn/skgz/bwyc/202410/t20241029_5797216.shtml (China).

[14]*See* David Adam, *Supportive? Addictive? Abusive? How AI Companions Affect Our Mental Health*, Nature (May 6, 2025), https://www.nature.com/articles/d41586-025-01349-9.

[15]*See* Lorenzo Brusci, *Humanity Reimagined: Toward an Anthropological Singularity*, LinkedIn (Apr. 20, 2024), https://www.linkedin.com/pulse/humanity-reimagined-toward-anthropological-lorenzo-brusci-iw12f/.

human beings demands fundamentally new legal and ethical frameworks to ensure the safe and beneficial development and use of AI systems. We cannot simply rely on the ethical norms and best practices developed for earlier technologies. Instead, we must directly grapple with the unique and weighty implications of artificial intelligence as it rapidly advances. The three areas highlighted by the author provide a framework for some of the key ethical fault lines that this will involve.

Considering these fundamental issues, this Article argues that AI value alignment should be adopted as a new approach to AI governance in China's AI industry, and even be a major focus of the global AI governance. The core objective is to make a case that beyond China's existing laws, regulations, and ethical guidelines, a focused emphasis on value alignment will enhance AI governance. By value alignment, I mean not only aligning AI with legal requirements, but ensuring AI systems inherently learn and behave in accordance with society's ethical values and the public interests. I posit that integrating value alignment into China's AI governance framework can: (1) Address current gaps in ensuring AI systems truly uphold ethical principles in practice; (2) better safeguard against emerging risks from advanced AI, such as future AGI, by preemptively aligning AI with human and societal values; and (3) reconcile technical design of AI with China's normative goals, for example, "human-centric" and socially beneficial AI, thereby strengthening public trust and legitimacy of AI.

To support this argument, the Article is structured as follows. Section B provides an overview of China's current AI governance landscape, examining its legal regulations, ethical principles, and industry self-regulation initiatives, and identifying strengths and limitations. Section C introduces the concept of AI value alignment in depth—its definition, importance for future AI governance, and the technical and normative challenges it entails, including for advanced AI systems. Section D discusses how principles of AI value alignment can be integrated into various facets of AI governance—from law and policy to ethical guidelines and corporate practices—to complement and reinforce China's existing AI governance framework. Finally, Section E concludes with the findings and suggests future directions for AI governance in China, highlighting how a value alignment approach could position China's AI industry on a path toward responsible and human-aligned AI development, and align its approach to AI governance with international AI governance initiatives.

By drawing on policy analysis and academic research, this Article aims to demonstrate that a value alignment approach offers a forward-looking evolution of AI governance—one that can help ensure China's powerful AI industry develops in accordance with the values of its people and the broader interests of humanity. In the long run, making AI value alignment a central governance principle in China could contribute not only to domestic goals of "safe, reliable, and controllable" AI, but also to global efforts to ensure AI remains beneficial and aligned with human well-being and interests, and is always under human's control.

B. China's AI Governance Landscape

In recent years, China has emerged as a global leader in artificial intelligence and is keenly aware that governing AI development and use is of paramount importance. Alongside its rapid advancement, policymakers recognize that while AI promises great economic and social benefits, it also poses risks—from privacy violations to algorithmic bias and even long-term safety concerns—that must be managed through effective governance.[16] In short, China views sound AI governance as integral to ensuring AI contributes to national development and social stability while minimizing harms.[17]

[16]Rebecca Arcesati, *Lofty Principles, Conflicting Incentives: AI Ethics and Governance in China*, MERICS (June 24, 2021), https://merics.org/en/report/lofty-principles-conflicting-incentives-ai-ethics-and-governance-china.

[17]*See* Hipolito Calero, *An Analysis of China's AI Governance Proposals*, CSET (Sep. 12, 2024), https://cset.georgetown.edu/article/an-analysis-of-chinas-ai-governance-proposals/#:~:text=AI%20as%20a%20%E2%80%9Cmajor%20strategic,put%20into%20effect%20in%202021.

China's approach to AI governance is multifaceted, involving top-down government regulations, high-level ethical guidelines, and contributions from industry. This part reviews the key components of this landscape: (I) Binding AI-related legislations and regulations promulgated by the government; (II) official AI ethics principles guiding AI development and use; (III) industry self-regulation efforts by major Chinese tech companies; and (IV) an overall assessment of the strengths and limitations of China's current AI governance regime. Understanding this status quo provides the foundation for why an additional focus on value alignment is needed.

I. AI Legislations

In recent years, China has enacted a series of laws and regulations that, while not always specific to AI, establish the legal parameters for AI development and use. At the national legislation level, three foundational laws form the pillars of China's digital governance framework: The Cybersecurity Law (CSL), effective 2017; the Data Security Law (DSL), effective 2021; and the Personal Information Protection Law (PIPL), effective 2021. Together, these laws create an overarching regime for cybersecurity, data governance, and privacy protection in China. Beyond these foundational laws, China has been quick to introduce AI-specific regulations, making it a pioneer in AI governance globally. Specifically, algorithmic recommendation systems, automated algorithmic decision-making, AI-driven deep synthesis, generative AI, and other algorithmic and AI applications within the internet sector have become key targets for government regulation in the AI domain.[18] This heightened regulatory attention is due to their extensive application, significant public interests, and persistent emergence of negative issues. Accordingly, China has introduced a series of laws and regulations aimed at actively standardizing and overseeing algorithmic and AI applications in the internet field.[19]

Regarding algorithmic recommendations, the Provisions on the Administration of Algorithmic Recommendations for Internet Information Services, which is issued by the Cyberspace Administration of China (CAC) and came into effect on March 1, 2022, constitute China's first legislation specifically focused on algorithmic governance.[20] These regulations comprehensively outline obligations and prohibitions for algorithmic recommendation service providers, establishing a range of regulatory measures such as algorithm security assessment and monitoring, algorithm registration management, and transparency requirements for algorithmic mechanisms, thereby significantly strengthening platform companies' responsibility for algorithmic safety and security. Industry insiders have marked the introduction of this regulation as the beginning of the era of algorithmic regulation, dubbing the year 2022 the "first year of algorithm regulation" in China.[21]

Regarding algorithmic automated decision-making, several regulations in China, including the PIPL, the E-Commerce Law, and the Interim Provisions on the Administration of Online Tourism

[18]*See* Angela Huyue Zhang, *The Promise and Perils of China's Regulation of Artificial Intelligence*, 63 COLUM. J. TRANSNAT'L L. 1 (2025).

[19]*See* Jet Deng & Ken Dai, *DeepSeek and China's AI Regulatory Landscape: Rules, Practice and Future Prospects*, JD SUPRA (Mar. 12, 2025), https://www.jdsupra.com/legalnews/deepseek-and-china-s-ai-regulatory-4601861/#:~:text=At%20the%20fundamental%20legal%20level%2C,synthesis%20service%20providers%20and%20technical.

[20]Hùliánwǎng xìnxī fúwù suànfǎ tuījiàn guǎnl guiding (互联网信息服务算法推荐管理规) [Provisions on the Administration of Algorithmic Recommendations for Internet Information Services] (promulgated by the Cyberspace Admin. of China, Ministry of Indus. & Info. Tech., Ministry of Pub. Sec., and State Admin. for Mkt. Regul., Nov. 16, 2021, effective Mar. 1, 2022), CYBERSPACE ADMIN. CHINA, Dec. 31, 2021, https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm (China).

[21]Yu Shengqi, *2022 Annual Review of Algorithmic Governance,* DIGITAL RULE OF L., https://mp.weixin.qq.com/s/c8tL5yNytTi6L-KyZBQbeQ (Nov. 5, 2023, 9:00PM) (China).

Business Services[22], contain provisions aimed at regulating unfair or improper algorithmic decision-making practices, such as algorithmic discrimination and differential pricing based on big data, often known as "big data-enabled price discrimination." The Provisions on Recommendation Algorithms also prevent abuses such as discriminatory pricing or the creation of addictive user feeds. These laws and regulations introduce regulatory measures such as granting individuals the right to choose or opt-out of automated decisions, requiring organizations to conduct impact assessments, and ensuring individuals have the right to request explanations and refuse algorithmic decision-making outcomes under certain circumstances. Collectively, these measures aim to better balance the protection of individual rights with the commercial applications of algorithms.

Regarding deep synthesis and generative AI, especially AI-generated contents, the CAC also introduced several relevant regulations. "Provisions on Deep Synthesis,"[23] effective January 2023, aimed at so-called deepfake technologies. These rules mandate that synthetically generated media—for example, AI-generated faces or voices, be clearly labeled to prevent misinformation, and they prohibit using deep synthesis in ways that could disrupt social order or infringe individuals' rights. In mid-2023, Chinese regulators released the Interim Measures for the Management of Generative AI Services[24]—the first national regulations on generative AI, such as large language models and image generators. These measures set requirements for generative AI providers to ensure the security, accuracy, and appropriateness of AI-generated contents, and they reinforce that generative AI must be value-aligned with China's laws and regulations, and respect social morality and ethics. In March 2025, Measures for the Labeling of AI-Generated Synthetic Content[25] was released by the CAC, which contains specific labeling requirements for different types of AI-generated contents. China is also actively developing standards to support these regulations. For example, authorities have issued the Basic Security Requirements for Generative AI, effective 2024, as a national standard, and are working on standards for labeling AI-generated content—including the standard "Methods for Labeling AI-Generated Synthetic Content."[26] All these efforts reflect an approach of "inclusive and prudent, classified and graded" regulation—meaning China tries to encourage AI innovation while keeping a close watch on high-risk uses.

In addition to national rules, local governments in China have issued AI-related regulations or guidance, such as provincial measures to encourage AI industry development, though these tend to focus on fostering industry rather than restrictions. A notable example is Shenzhen's Regulations on the Promotion of the Artificial Intelligence Industry. As a Special Economic Zone, Shenzhen has been granted special legislative powers by the National People's Congress, which means that Shenzhen can explore legislative innovations in fields such as autonomous driving and artificial intelligence. In addition to measures supporting the development of the AI industry,

[22]Zàixiàn l yóu jīngyíng fúwù guǎnl zhàn háng guiding (在线旅游经营服务管理暂行规定) [Interim Provisions on the Administration of Online Tourism Business Services] (promulgated by the Ministry of Culture and Tourism, July 20, 2020, effective Oct. 1 2020), Ministry Culture & Tourism, Aug. 20, 2020, https://www.gov.cn/zhengce/zhengceku/2020-09/01/content_5538951.htm (China).

[23]Hùliánwǎng xìnxī fúwù shēndù héchéng guǎnl guiding (互联网信息服务深度合成管理规定) [Provisions on the Administration of Deep Synthesis of Internet Information Services] (promulgated by the Cyberspace Admin. of China, Ministry of Indus. & Info. Tech., and Ministry of Pub. Sec. Nov. 3, 2022, effective Jan. 10, 2023), Cyberspace Admin. China, Nov. 25, 2022, https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm (China).

[24]*See* Shēngchéng shì réngōng zhìnéng fúwù guǎnl zhànxíng bànfǎ (生成式人工智能服务管理暂行办法) [Interim Measures for the Management of Generative AI Services] (promulgated by the Cyber Space Admin. of China May 23, 2023, effective Aug. 15, 2023), China Internet Info. Off., July 13, 2023, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm (China).

[25]*See* Réngōng zhìnéng shēngchéng héchéng nèiróng biāozhì bànfǎ (人工智能生成合成内容标识办法) [Measures for the Labeling of AI-Generated Synthetic Content] (promulgated by the Cyberspace Admin. of China Mar. 7, 2025, effective Sept. 1, 2025), China Internet Info. Off., Mar. 14, 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm (China).

[26]*Cybersecurity Technology—Labeling Method for Content Generated by Artificial Intelligence*, Cyberspace Admin. China, Mar. 15, 2025, https://www.cac.gov.cn/2025-03/15/c_1743755846973028.htm (China).

Shenzhen's AI regulations also emphasize AI ethics and governance, requiring that organizations or individuals engaged in AI research and applications shall comply with relevant ethical and safety requirements for artificial intelligence, and conduct ethical and safety reviews and risk assessments concerning possible adverse impacts of their AI products and services on national interests, public safety, commercial order, and individual rights and interests.

In sum, China's legislative landscape for AI governance is characterized by a top-down, comprehensive strategy. Fundamental laws (CSL, DSL, PIPL) create broad obligations for security and privacy that AI developers must heed. Meanwhile, specific regulations target AI-related risks—from algorithms' social impacts to deepfake abuses—often prescribing that AI systems align with legal and societal norms.[27] China's early move to regulate algorithms in 2022 demonstrates a willingness to be at the forefront of AI governance globally. However, these laws and regulations primarily address external governance—what AI systems are permitted to do, data they can use, content they must avoid—rather than the internal mechanisms by which AI aligns with ethical values and norms through alignment training. This opens the question of whether a more intrinsic approach—such as building ethical values into AI systems through alignment training and other techniques—could complement the existing governance framework.

II. AI Ethics

Ethical considerations form a cornerstone of China's approach to AI governance, with formal frameworks establishing principles for responsible development and deployment.[28] The New Generation AI Development Plan, 2017, marked ethics as a key component of China's AI strategy. In recent years, China has continuously worked to establish and improve its tech ethics framework, fostering a culture of responsible and ethical technology and building safeguard mechanisms to ensure that technology serves the public good, in other words, a culture of tech for good. For example, a major highlight of the revised Law on Progress of Science and Technology, amended in December 2021, is the inclusion of provisions related to science and technology ethics. On one hand, the law establishes general regulations for improving the governance system of science and technology ethics; on the other hand, it mandates that research and development institutions, universities, enterprises, and public institutions assume primary responsibility for tech ethics management, conduct ethical reviews of scientific and technological activities, and adhere to the legal and ethical bottom lines in research, development, and application. The policy document titled *Opinions on Strengthening the Governance of Science and Technology Ethics*[29] proposes a more comprehensive framework for the governance of tech ethics, covering six key areas: Overall requirements, guiding principles, governance structures, institutional safeguards, review and supervision mechanisms, and education and public awareness. This provides a solid foundation for the practical implementation of science and technology ethics governance. Current policies and regulations seek to further define the primary responsibilities for science and technology ethics management, requiring innovation entities to establish ethics committees, uphold the ethical bottom lines of science and technology, conduct ethics reviews, implement risk monitoring, early warning, and assessment mechanisms, and carry out ethics training. As we can see from these laws and policies, China takes an ethics-first approach to AI innovation, requiring

[27] *See* Matt Sheehan, *China's New AI Governance Initiatives Shouldn't Be Ignored,* CARNEGIE ENDOWMENT FOR INT'L PEACE (Jan. 4, 2022), https://carnegieendowment.org/posts/2022/01/chinas-new-ai-governance-initiatives-shouldnt-be-ignored?lang=en.

[28] *See* Samuel Yang, Chris Fung & Bill Zhou, *China Law Vision: AI Ethics: Overview (China)*, ANJIE BROAD (Jan. 20, 2025), https://www.chinalawvision.com/2025/01/digital-economy-ai/ai-ethics-overview-china/.

[29] *The General Office of the CPC Central Committee and the General Office of the State Council issued the "Opinions on Strengthening the Governance of Science and Technology Ethics"*, XINHUA NEWS AGENCY (Mar. 20, 2022, 7:05PM), https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm (China).

imbedding ethical principles and requirements into the life cycle of AI activity, which is aligned with the notion of responsible AI or trustworthy AI.

Under this overall tech ethics framework, Chinese authorities have promulgated ethical guidelines and principles to steer AI development in a socially desirable direction. China's leadership stresses a "human-centric" and "ethical" development of AI, and in recent years several high-level ethical frameworks have been issued. For example, National Governance Committee for New Generation AI under China's Ministry of Science and Technology (MOST) has successively released the documents: *Governance Principles for New Generation of AI—Developing Responsible AI,*[30]2019, and *Ethical Norms for New Generation AI,*[31] 2021, providing ethical guidance for the development of responsible AI.

These ethical guidelines are advisory in nature—they are not laws with penalties, but they carry normative weight and signal the government's expectations. They also serve as a basis for developing standards and for guiding companies' internal policies. One strength of China's approach is the high-level consistency: Documents from 2019 to 2021 reiterate themes of human-centric, fair, secure and reliable, beneficial, and controllable AI. The term "controllable" is often stressed, reflecting China's concern that AI systems remain under human oversight and control. In practice, these principles set an ethical tone: For example, developers are encouraged to design AI that improves livelihoods and avoids worsening inequality.

It is worth noting, however, that these ethics policies lack formal enforcement mechanisms. They rely on voluntary compliance and integration into other policy tools—standards, certifications, and so forth. This gap between principle and practice is a recurring theme we will address. Nonetheless, China's ethical guidelines provide a conceptual foundation for value alignment, because they articulate the values AI is expected to uphold—human welfare, fairness, and others. The next challenge is ensuring AI systems actually behave in accordance with these values—which is precisely what the value alignment approach seeks to tackle.

III. AI Industry Self-Regulations

Effective AI governance requires the participation of multiple stakeholders, including the government, enterprises, industry organizations, academic communities, users or consumers, and the broader public. Among these, self-regulation and internal governance by tech companies are crucial means of implementing the principle of "ethics first." The call for "ethics first" arises largely in response to the increasingly apparent lag of legal systems in keeping up with the rapid development of AI. In the AI domain, "ethics first" is primarily embodied in corporate self-regulation of science and technology ethics, with best practices from leading companies often playing a demonstrative and guiding role, helping to drive the entire industry toward responsible innovation and ethical AI development.

Therefore, China's AI governance landscape also features significant involvement from the private sector, including industry self-regulatory initiatives and internal corporate governance of AI. The government actively encourages companies to follow ethical principles and sometimes facilitates industry pledges. Meanwhile, the industry is actively exploring self-regulatory measures for AI governance, practicing the principles of responsible innovation and tech for good. Leading tech companies in China have begun establishing their own AI ethics committees and policies. These efforts can be seen as steps toward self-regulation, complementing government rules.

At the industry level, in recent years, relevant research institutions and industry associations have issued various ethical guidelines and self-regulatory codes to provide ethical direction for

[30] *Developing Responsible AI: Release of New-Generation AI Governance Principles*, Ministry Sci. & Tech. (Jun. 17, 2019), https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html.

[31] *"Ethical Norms for New Generation AI" Released*, Ministry Sci. &Tech. (Sep. 26, 2021), https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

corporate AI activities. Notable examples include: (1) The *AI Industry Self-Discipline Convention*[32], the *AI Safety Commitments*[33] and the *Trustworthy AI Operational Guidelines* by the Artificial Intelligence Industry Alliance (AIIA);[34] (2) The *Beijing AI Principles*[35] and the Declaration of *AI Industry Responsibility* by the Beijing Academy of Artificial Intelligence (BAAI).[36]

These voluntary agreements brought together major AI players to agree on voluntarily adhering to principles like fairness, transparency, user safety, and avoiding misuse of AI. They underscore a trend where industry participates in shaping AI governance under regulatory authorities' endorsement. They focus on "applicable and action-oriented goals" to ensure AI development is beneficial to society across the lifecycle, from research and development to deployment.

At the corporate level, as the key drivers of AI innovation and application, technology companies bear significant responsibility for the responsible research, development, and deployment of AI. Internationally, technology firms have developed relatively mature practices, ranging from proposing AI ethics principles, establishing internal AI ethics governance bodies, to developing management tools, technical solutions, and even commercial offerings for responsible AI—many of which provide transferable and scalable experiences.[37] In recent years, China's tech giants have actively responded to regulatory demands, and started to institute their own AI governance structures and explored self-regulatory practices, including:

(1) Publishing AI ethics principles;
(2) Establishing internal AI governance structures, such as ethics committees or ethics review board;
(3) Conducting ethical reviews or safety risk assessments of AI activities;
(4) Publishing transparency reports or writing articles on AI ethics, and disclosing algorithm-related information to promote algorithmic transparency;
(5) Exploring technical solutions to AI ethics challenges—for example, detection tools for synthetic content, AI governance evaluation tools, and privacy-preserving computation methods such as federated learning.

Industry self-regulation in China often happens in close coordination with government expectations. The state's push for "social responsibility" among tech firms means companies have an incentive to preempt stricter regulation by self-policing their AI products. It also appears in how companies publicize alignment with national initiatives—like using AI for public good projects, or emphasizing "AI for Social Good" in corporate messaging—aligning corporate values with national values. However, how far self-regulation goes in practice can vary. Companies still face competitive pressures and may not sacrifice profit for ethics unless compelled or unless it also enhances user trust.

[32]Graham Webster, *Translation: Chinese AI Alliance Drafts Self-Discipline "Joint Pledge"*, New America (June 17, 2019), https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/.

[33]Scott Singer, *DeepSeek and Other Chinese Firms Converge with Western Companies on AI Promises*, Carnegie Endowment for Int'l Peace (Jan. 28, 2025), https://carnegieendowment.org/research/2025/01/deepseek-and-other-chinese-firms-converge-with-western-companies-on-ai-promises?lang=en.

[34]*AIIA Unveils "Trustworthy AI Operational Guidelines" Version 0.5*, AIIA (Aug. 18, 2020), https://mp.weixin.qq.com/s/RB6Y8gvow3mVIsbYWYErfw (China).

[35]Beijing Academy of Artificial Intelligence, *Beijing AI Principles*, 10 Datenschutz und Datensicherheit 656, 656 (2019), https://link.springer.com/content/pdf/10.1007/s11623-019-1183-6.pdf.

[36]*Leading AI Enterprises Release First "Declaration of AI Industry Responsibility,"* Guancha (Aug. 4, 2021), https://www.guancha.cn/industry-science/2021_08_04_601565_3.shtml (China).

[37]Cao Jianfeng & Hu Jinhao, *Ethics as a Service: The Next Wave of Tech Ethics and Trustworthy AI*, Tencent Rsch. Inst. (Aug. 17, 2021, 6:00 AM), https://mp.weixin.qq.com/s/nBqgBjmwz1RIyXH-BhwclA (China).

In summary, Chinese AI companies are increasingly engaging in self-governance of AI, through pledges, ethics committees, and internal principles that often parallel national AI ethics principles and guidelines. This trend indicates an understanding that long-term business sustainability in AI requires addressing ethical and safety risks, and aligning with societal values. Such industry efforts provide a foundation upon which a more explicit value alignment approach could build, as we will discuss, because they reflect a growing commitment to ensure AI does what is socially acceptable and beneficial.

IV. Remarks on China's AI Governance

China's current AI governance system is notable for its proactiveness and breadth.[38] Few countries have as comprehensive a suite of AI-related regulations and official ethics guidelines as China does as of the early 2020s. China moved early to regulate areas like recommendation algorithms and deepfakes (deep synthesis), showcasing a willingness to intervene in industry practices for societal interests. There is a strong emphasis on ethical principles—human-centric, AI for good, fairness, safety, beneficial, and so on—in strategy documents, and these principles are increasingly being codified into both soft guidelines and hard rules. Another strength is the multi-stakeholder element: While the government ultimately drives policy, it has involved experts, companies, and academics in drafting AI-related regulations, guidelines, and standards. This inclusive approach has helped produce well-rounded principles that consider technical feasibility and industry perspective, not just abstract regulatory ideals. Moreover, China's framework combines different instruments—law, standards, ethics codes, and industry commitments—which together can cover both general and specific issues. The use of standards and pilot programs, for example, AI governance experimental zones, allows testing and refining governance approaches in practice. In terms of enforcement capacity, China's centralized governance model means regulations, for example, the algorithm filing requirement, can be implemented relatively quickly through its bureaucratic apparatus.

Despite these strengths, there are several limitations and challenges in China's AI governance that motivate the exploration of new approaches like value alignment.

1. Principle-Practice Gap

A recurring challenge is ensuring that lofty principles translate into actual practice. Many of AI ethics guidelines are broad and lack enforcement mechanisms. For example, an AI company can pledge to be fair and "people-centered," but measuring and guaranteeing this in their algorithms is difficult. The governance system currently relies heavily on after-the-fact supervision and enforcement, for example, punishments if content violations are found or AI systems cause social harms, and companies' self-declaration of compliance. There is less in the way of systematic assessments and audits or technical verification that an AI system is aligning with ethical norms. This gap leaves room for non-aligned outcomes, such as algorithms still exhibiting unlawful bias or making opaque decisions contrary to user interests, even if unintentionally.

2. Regulatory Fragmentation and Adaptability

While China has many AI regulations, they have emerged in a somewhat piecemeal fashion (each addressing a specific technology or issue). There isn't yet a single comprehensive "AI Law" akin to the European Union's (EU) enacted AI Act, though legal scholars have proposed drafts for one. This fragmentation can lead to overlaps or gaps—for example, if a new AI application doesn't squarely fit existing rules. Furthermore, AI technology evolves quickly, and rule-making can lag

[38] *See* Matt Sheehan, *China's AI Regulations and How They Get Made*, Carnegie Endowment for Int'l Peace (July 10, 2023), https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en.

behind; the challenge is keeping regulations updated without stifling innovation. The current approach, which is often reactive—for example, deepfake rules after deepfakes became prominent—might struggle as AI systems become more complex in design and more general and autonomous in capabilities—such as, how to regulate an autonomous AI agent that doesn't fit a single domain. In this sense, focusing on aligning AI systems with ethical principles and values from the design stage could be a more proactive, flexible, long-term solution than always chasing new rules for new applications.

3. Corporate Compliance vs. Initiative

Industry self-regulation, while present, may often be driven by compliance mentality—"tick-box" adherence to government demands—rather than proactive internalization of ethical values. Companies might publish ethics charters due to external pressure, but the true test is in product design decisions and investment trade-offs. If ethics is siloed as a public relations or policy matter, it may not deeply influence engineering. Some analyses have pointed out that Chinese AI companies have produced relatively fewer research publications on AI safety or alignment compared to their Western counterparts.[39] This suggests that, so far, much of the serious technical work on ensuring AI safety, including ethical alignment, in China is happening in academia or research labs, not as much within industry. Strengthening the mandate or incentives for companies to build value alignment into their AI research and development could address this.

4. Transparency and Trust

Another limitation in the current governance model is a relative lack of transparency. Many of the AI oversight processes, like algorithm filings or security assessments, happen behind closed doors between companies and regulators. The public often has limited insight or input. This can undermine trust—users might not know whether an AI system is ethical and truly respecting their rights or just following minimum legal requirements. An alignment-focused governance might push for more transparency about how AI system aligns with ethical principles and values—for example, publishing model specifications (information concerning AI model's behavioral boundary, design principles, value norms, and so on), or showing users why an AI system made a decision—which would improve accountability.

In light of the above, China's AI governance is strong in setting external requirements and desired end-states—goals and outcomes—for AI systems, such as, "AI must not violate privacy, must be fair and transparent, etc." But it faces challenges in ensuring AI systems internally adhere to those values when taking actions or interacting with other actors in complex, real-world scenarios. This is where AI value alignment can contribute: It shifts some focus to the intrinsic design of AI systems so that they are by design aligned with human values and norms, reducing the burden on after-the-fact enforcement. In the next part, I delve into what AI value alignment means and why it is becoming vital for the future development of AI governance, both in China and globally.

C. AI Value Alignment as a New Approach to AI Governance

In this part, I turn to the concept of AI value alignment and explore why it represents a crucial frontier for AI governance. I first explain the concept and its theoretical foundations. Second, I discuss why value alignment is increasingly seen as vital to managing the future development of AI—particularly as AI systems grow more advanced and autonomous—and how this relevance applies in the Generative AI context. Finally, I examine the core issues and challenges that arise in

[39]*See* Matt Sheehan, *China's Views on AI Safety Are Changing—Quickly*, Carnegie Endowment for Int'l Peace (Aug. 27, 2024), https://carnegieendowment.org/research/2024/08/china-artificial-intelligence-ai-safety-regulation?lang=en.

pursuing AI value alignment, dividing them into (1) technical aspects, (2) normative (value) aspects, and (3) specific considerations for advanced AI such as AGI or superintelligence.

I. The Concept of AI Value Alignment

In the field of AI, AI value alignment generally refers to the process of ensuring that an AI system's goals, intentions, and behaviors are aligned with the values, ethics, and intentions of humans. In other words, this process of value alignment aims to ensure that AI acts in ways beneficial to humanity and society at large, without causing harm or disruption to human values and rights, even inadvertently.[40] As AI systems become more autonomous and powerful, ensuring they remain "properly aligned with human values" and "amenable to human control" is seen as vital.[41] A simple analogy might help illustrate this concept; imagine you have a highly intelligent dog. You might train it to fetch the newspaper every morning. However, if your instructions aren't careful enough, problems can occur—perhaps the dog might pick up all your neighbors' newspapers or shred the paper it retrieves. Designing AI systems can be similar: You want the AI to perform specific tasks, but you must provide careful, precise instructions, or unintended issues could arise.

Early computing pioneer Norbert Wiener foreshadowed this challenge in 1960 when he warned that if we delegate decisions to machines, "we had better be quite sure that the purpose put into the machine is the purpose which we really desire."[42] In the 2000s and 2010s—as AI, especially machine learning and deep learning, advanced—thinkers like Eliezer Yudkowsky and Nick Bostrom began warning about the "alignment problem" in the context of superintelligent AI. If a superhuman AI's values aren't aligned, it could cause catastrophic outcomes.[43] By the mid-2010s, leading AI institutes—such as MIRI, OpenAI, and DeepMind[44]—devoted research to alignment, and the term "AI alignment" became widely used. More recently, AI scholar Stuart Russell described misalignment as when we "inadvertently imbue machines with objectives that are imperfectly aligned with our own."[45] This misalignment could lead to AI pursuing a course of action that optimizes for the wrong objective—for example, a hypothetical housekeeping robot that is told to "make the house clean" might throw away everything including items of personal value, because it wasn't aligned with the human's actual intent. Another example would be a hypothetical childcare robot that is told to "make sure the child is well-fed" might, upon finding the refrigerator empty, decide to kill the family pet and cook it for the child.

Importantly, alignment is not just about preventing apocalyptic scenarios; it's also about everyday AI ethics. A 2021 article by Iason Gabriel frames value alignment as a continuum from near-term issues, like fair algorithms, to long-term existential safety.[46] In everyday terms, alignment includes making sure a content recommendation algorithm aligns with values like not promoting misinformation or hate speech, or that an autonomous vehicle respects the value of human life in how it makes split-second decisions.

---

[40] *See* Benjamin Larsen & Virginia Dignum, *AI Value Alignment: How We Can Align Artificial Intelligence with Human Values*, World Econ. F. (Oct. 17, 2024), https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/.

[41] Iason Gabriel & Vafa Ghaazavi, *The Challenge of Value Alignment: from Fairer Algorithms to AI Safety*, *in* The Oxford Handbook of Digit. Ethics (forthcoming 2022) (manuscript at 1), https://arxiv.org/pdf/2101.06060#:~:text=the%20prominent%20AI%20researcher%20Stuart,2019%2C%20137.

[42] Norbert Wiener, *Some Moral and Technical Consequences of Automation*, 131 Sci., New Series 1355, 1355–58 (1960), https://www.cs.umd.edu/users/gasarch/BLOGPAPERS/moral.pdf.

[43] *See* Ben Pace, *The Failed Strategy of Artificial Intelligence Doomers*, LessWrong (Jan. 31, 2025), https://www.lesswrong.com/posts/YqrAoCzNytYWtnsAx/the-failed-strategy-of-artificial-intelligence-doomers.

[44] *See* Iason Gabriel, *Artificial Intelligence, Values and Alignment*, Google DeepMind (Jan. 13, 2020), https://deepmind.google/discover/blog/artificial-intelligence-values-and-alignment/.

[45] Stuart Russell, Human-Compatible Artificial Intelligence 3, https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf.

[46] *See* Gabriel & Ghazavi, *supra* note 41, at 1–2.

In the field of generative AI and frontier AI models, as large language models continue to grow in capabilities and become increasingly generalized, ensuring that the behaviors and objectives of these models align with human's values, preferences, ethics, intentions, and goals has emerged as a crucial focus for their development.[47] Because a large language model (LLM) without value alignment could produce content containing racial or gender discrimination,[48] assist cybercriminals by generating code or content used in cyberattacks and telecommunications fraud, attempt to persuade or assist suicidal users in ending their lives,[49] or otherwise create harmful outputs. For example, OpenAI's early language model was found to output racist or biased language when prompted in a certain way; this was a failure of alignment with broader societal values of respect and fairness.

In summary, AI value alignment is a relatively new concept in the fields of AI safety and ethics, whose primary objective is to shape large AI models into safe, honest, useful, and harmless intelligent assistants, thus preventing potential undesired outcomes, negative consequences or harms during human interactions—such as generating harmful contents, producing hallucinations, or creating discriminatory outputs. Furthermore, AI value alignment as a concept emerged from the recognition that technical solutions are needed to embed ethical and value constraints into AI behavior. It is not sufficient to rely on external rules or hope that AI developers simply don't make mistakes—alignment calls for intentional design that makes AI models inherently follow human values and preferences. As we move forward, this notion of building AI that "does the right thing or takes moral action" by design, the so-called AI ethics by design, is increasingly seen as an essential complement to traditional governance. Therefore, to achieve AI ethics by design, value alignment is a vital approach.

II. Why AI Value Alignment is Vital to Future AI Governance

In the author's view, AI value alignment is important not only because it helps prevent harm and ensures ethical behavior, but also because it is a necessary response to the growing autonomy and unpredictability of advanced AI systems. As traditional governance mechanisms and tools struggle to keep pace, value alignment offers a scalable, forward-looking solution that bridges the gap between policy goals and algorithmic behavior. It is not just a technical issue—it is a governance imperative for the age of frontier AI.

1. Preventing Accidents and Unintended Consequences

The first reason value alignment is vital is to prevent the myriad unintended harms that misaligned AI can cause. Even today's relatively narrow AI systems have demonstrated problems when not properly aligned with human values. For instance, machine learning algorithms in lending or hiring have sometimes learned biased decision rules that discriminate against certain groups, because the algorithms were optimizing for accuracy or profit without a value constraint of fairness. One could imagine an AI-driven decision-making system for social services needing alignment to avoid marginalizing rural populations or the poor. Misaligned content recommendation systems have been implicated in spreading misinformation or extremist content on social media—outcomes at odds with legal requirements or social morality. If AI systems are left to optimize purely on short-term user engagement, they may sacrifice truth or

[47]Cao Jianfeng, *Human-AI Alignment: The Inevitable Path to Artificial General Intelligence*, TENCENT RSCH. INST. (Nov. 1, 2024, 3:00 AM), https://mp.weixin.qq.com/s/iOOapnk9pbC8W5dHR5ALpA (China).

[48]*See* Katharine Miller, *Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes*, STAN. UNIV. (Sept. 3, 2024), https://hai.stanford.edu/news/covert-racism-ai-how-language-models-are-reinforcing-outdated-stereotypes.

[49]*See* Chloe Xiang, *'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says*, VICE (Mar. 30, 2023), https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/.

social cohesion. Aligning these systems with values like truthfulness, or with relevant social norms and values, would help mitigate such harms, and make sure AI systems act predictably and safely.

AI value alignment is especially vital for large language models and other generative AI systems, because it has the potential to address many of the current safety and ethical challenges associated with generative AI. These challenges include harmful outputs, hallucinations and confabulations, jailbreaks, and other adversarial attacks. Additionally, value alignment can help mitigate issues like discrimination and bias, regurgitation of copyrighted information, and other emergent risks and catastrophic risks—such as power-seeking, deception, undue persuasion, cybersecurity threats, biological threats (Chemical, Biological, Radiological, and Nuclear), and concerns related to the autonomy of AI models—such as autonomous AI research and development, self-replication, shutdown avoidance, and circumvention of oversight mechanisms.[50] By aligning AI systems with human values and ethical principles, we can work towards a safer and more ethical deployment of generative AI technologies.

2. Scaling with AI Capabilities

As AI capabilities grow, the importance of alignment grows exponentially. We are entering an era of increasingly accelerating "frontier AI"—highly general, powerful models, like GPT-4, that can perform a wide array of tasks and even develop novel behaviors. These systems are more difficult to predict and constrain. Traditional governance mechanisms (laws, checklists, ethical reviews) might not be fast or fine-grained enough to manage such systems' behavior in real time. If a model can make autonomous decisions—for example, an AI agent might autonomously trade stocks, manage infrastructure, or interact with humans conversationally at scale—we need it to fundamentally share our values and moralities, because constant human monitoring of every action becomes infeasible. In China's context, the government has explicitly recognized AI safety as a rising concern: Recent discourse shows Chinese scientists and policymakers discussing how to ensure advanced AI remains safe. A major policy document in 2024 even called for "oversight systems to ensure the safety of artificial intelligence," indicating awareness at the highest levels that AI's development trajectory needs safety guidance.[51] By investing in value alignment research and implementation, China can better future-proof its AI governance against the challenges posed by increasingly autonomous AI systems.

3. Alignment with Societal Values

AI does not exist in a vacuum—it operates within societies and can influence or even shape social values.[52] Therefore, ensuring AI aligns with societal values and building ethical AI is a governance imperative. Aligned AI systems are more likely to operate within ethical boundaries, minimizing harm and ensuring fairness. Conversely, if AI systems were seen as eroding cherished values or threatening social norms, public backlash could ensue. In a more positive framing, value alignment is vital to maximize AI's benefits for society. When AI systems are aligned with what people truly value, they can become more trustworthy and useful, and can be more effectively used as tools to enhance human wellbeing and societal welfare.

4. Addressing Limitations of Current Governance

We saw in Section B that one limitation of current governance is enforcement and the gap between high-level principles and algorithmic behavior. A value alignment approach is essentially a way to

[50] *See OpenAI o1 System Card*, OPEN AI (Dec. 5, 2024), https://openai.com/index/openai-o1-system-card/.

[51] *Why Establish an AI Safety Regulatory Framework*, XINHUANET (Nov. 6, 2024), https://www.qstheory.cn/qshyjx/2024-11/06/c_1130216759.htm (China).

[52] *See generally* Henry Farrell, Alison Gopnik, Cosma Shalizi & James Evans, *Large AI Models are Cultural and Social Technologies,* 387 SCI. 1153 (2025).

bridge that gap. It aims to bake the compliance in at the AI's design phase. For regulators, this could reduce the need for constant micromanagement and allow more focus on verifying alignment processes than chasing individual violations. For companies, achieving alignment would reduce the risk of violations or public relations disasters and likely become a competitive advantage, as users and clients prefer AI they can trust. Moreover, alignment provides a common language between technical and policy domains. It links the policy intent (ethical principle) with the technical implementation (AI's objective function). In China, where the government often sets broad goals—for example, "AI must be controllable and trustworthy," or "AI must be safe, fair, and beneficial"—alignment research can translate that into engineering terms, like how to ensure an AI always defers to a human override or always rejects requests to do unethical or unlawful tasks.

In summary, AI value alignment is increasingly important part of international AI governance conversations, and it is vital because it addresses the root cause of many AI risks—the misalignment between what we want AI to do and what it will try to do on its own. Actually, misalignment manifesting as instrumentally convergent subgoals arises in service of achieving predetermined objectives. For China, incorporating value alignment into AI governance regime not only supports its emphasis on proactive risk management and its vision of "human-centered AI, and AI for good,"[53] but also enhances its ability to coordinate with emerging international frameworks for AI governance. All in all, AI value alignment is a forward-thinking approach that complements existing regulatory and governance mechanisms by ensuring that as AI grows more capable, it remains tethered to the values and objectives that society chooses.

III. Core issues of AI Value Alignment

To achieve value alignment, AI developers need to ensure, at the model level, that AI models understand and adhere to human values, preferences, and ethical principles, thereby minimizing harmful outputs, misuse, misalignment, and mistakes. However, implementing AI value alignment is not a trivial task—it raises several complex issues. We can categorize these into technical challenges (how to do it in practice), normative questions (which values or ethical principles to align with and who decides), and special considerations for future advanced and powerful AI systems (in other words, AGI and superintelligence). The normative and technical dimensions of the AI value alignment problem are closely interconnected, creating opportunities for productive collaboration and dialogue between practitioners in both domains.[54]

1. Technical Aspects of Alignment: Two Main Approaches and Several Difficulties

From a technical perspective, AI value alignment is a crucial aspect of the development and training of large language models. Currently, mainstream AI enterprises are actively exploring how to achieve effective alignment and control of frontier AI models and even the so-called superintelligence. Technically, there are currently two main approaches to alignment:

1.1. The Bottom-Up Approach: Reinforcement Learning from Human Feedback (RLHF)

This method aligns AI systems through reinforcement learning based on human feedback.[55] It involves fine-tuning the model using datasets curated for value alignment, and having human trainers evaluate model outputs.[56] The feedback is then used to guide the model, helping it learn

[53]*Global AI Governance Initiative*, Ministry of Foreign Affs. People's Republic of China (Oct. 20, 2023), https://www.fmprc.gov.cn/eng/xw/zyxw/202405/t20240530_11332389.html.

[54]*See* Iason Gabriel, *Artificial Intelligence, Values and Alignment*, 30 Minds & Machines 411 (2020).

[55]Aravind Kolli, *Reinforcement Learning from Human Feedback (RLHF): An End-to-End Overview*, Medium (Feb. 11, 2024), https://aravindkolli.medium.com/reinforcement-learning-from-human-feedback-rlhf-an-end-to-end-overview-f0a6b0f06991.

[56]Kolli, *supra* note 55.

human values and preferences through reinforcement learning.[57] Technically, RLHF includes several key steps: Training an initial model, collecting human feedback, applying reinforcement learning, and undergoing iterative refinement.[58]

1.2. The Top-Down Approach: Principles-Based AI Alignment

This method centers on instilling a set of ethical principles into the model and using technical methods to enable the model to evaluate or score its own outputs in light of those principles. The aim is to ensure that the model's outputs conform to predefined ethical standards. Unlike OpenAI who adopts the RLHF-based bottom-up approach to alignment, the American AI company Anthropic employs the top-down, principle-based approach known as "Constitutional AI."[59] Specifically, this method involves developing a subordinate AI model whose primary function is to evaluate whether the outputs of a main model adhere to a set of predefined "constitutional" principles—that is, a set of guiding rules or norms established in advance.[60] The evaluation results generated by the subordinate model are then used to optimize the behavior of the main model.[61] Drawing on its own practical experience, Anthropic compiled a comprehensive list of principles inspired by documents such as the Universal Declaration of Human Rights, Apple's Terms of Service, and DeepMind's Sparrow rules.[62] These principles serve as the evaluative standard by which its large language model, Claude, assesses its own outputs, aiming to cultivate good values during training.[63] The goal is to maximize the helpfulness of the model's responses while minimizing the likelihood of generating harmful content.[64]

However, there is no unified technical roadmap for AI value alignment. Some researchers' analysis suggests that RLHF or Constitutional AI, as a standalone solution, may be insufficient to address all the challenges of AI value alignment. From a technical standpoint, aligning AI with human values is challenging because of the way modern AI systems learn and make decisions. Most cutting-edge AI uses machine learning algorithms, like deep neural networks, that learn patterns from data rather than being explicitly programmed. Ensuring such systems adhere to abstract values involves several difficulties:

1.3. Specifying the Right Objective

In reinforcement learning and other AI paradigms, designers specify a reward function or objective for the AI to optimize. One core alignment problem is that it's hard to formally specify what we really want in all cases. If the specification is even slightly off, a capable AI might exploit that loophole—a phenomenon known as reward hacking. For example, if an AI is rewarded for cleaning a room, it might push dirt under the rug (cheating) because the designers didn't explicitly penalize hidden dirt. Aligning AI requires methods and techniques to avoid such loopholes. Researchers have proposed solutions like iterative refinement of reward, to adjust the objective as you see unintended behaviors, or designing AIs that maintain uncertainty about the true objective and continuously ask for human guidance.[65]

---

[57] *Id.*

[58] *Id.*

[59] *See generally* Gilad Abiri, *Public Constitutional AI*, 59 GA. L. REV. 601 (2025).

[60] *Claude's Constitution*, ANTHROPIC (May 9, 2023), https://www.anthropic.com/news/claudes-constitution.

[61] *Id.*

[62] *Id.*

[63] *Claude's New Constitution*, ANTHROPIC (Jan. 22, 2026), https://www.anthropic.com/news/claude-new-constitution.

[64] *See* YUNTAO BAI ET AL., CONSTITUTIONAL AI: HARMLESSNESS FROM AI FEEDBACK 1 (Dec. 15, 2022), https://arxiv.org/pdf/2212.08073.

[65] *See* STUART RUSSELL, PROVABLY BENEFICIAL ARTIFICIAL INTELLIGENCE 7 (2025), https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf.

1.4. Learning Human Preferences

Another approach to value alignment is to have AI models learn values and preferences from human behavior or feedback, rather than rely only on a hand-written objective. Methods like inverse reinforcement learning (IRL) or preference learning allow an AI model to observe human decisions and infer what the humans value. While promising, these methods face issues: Humans are not perfectly rational and have inconsistent preferences. Also, AI models might misinterpret demonstrations. Moreover, they require that the human training signals, like ratings of AI outputs, are truly reflective of values, which can be noisy or biased. Ensuring AI models generalizes the learned values correctly to new situations is an ongoing research problem. As a main method to align current LLMs with human values, RLHF has been a key contributor to the success of ChatGPT, playing a central role in enabling the model to produce responses that are, to a significant extent, helpful, trustworthy, and harmless. However, despite its effectiveness, RLHF is not without limitations. One major concern is its limited scalability—the reliance on human annotators for feedback makes it labor-intensive and difficult to apply efficiently at a large scale. Additionally, the alignment achieved through RLHF is inherently constrained by the subjective preferences and potential biases of the human trainers, which may not reflect the full diversity of societal values. Perhaps more fundamentally, RLHF struggles to ensure alignment with long-term or abstract human values, especially in complex or evolving contexts where short-term preferences may not capture deeper ethical or normative considerations. These challenges highlight the need for complementary approaches and ongoing research in the field of AI alignment. For example, OpenAI is exploring a new alignment method called Deliberate Alignment, which addresses key limitations of traditional alignment techniques like RLHF by embedding policy comprehension and ethical reasoning into the model's decision-making process.[66]

1.5. Evaluating Alignment Effectiveness

Evaluating the effectiveness of alignment in LLMs remains a highly challenging task. Several factors contribute to this difficulty: First, standards vary across countries, regions, and organizations, each defining "trustworthiness" or "ethical AI" differently, making it difficult to reach a universal consensus. Second, the complexity of the tasks involved in trustworthy AI evaluation spans multiple dimensions—such as fairness, robustness, safety, and ethics. When layered across diverse real-world scenarios, no single set of tasks or metrics can comprehensively assess alignment performance. Third, there is a lack of data and tools. Compared to capability evaluations, there are fewer datasets and tools available for evaluating alignment and trustworthiness. The 2025 AI Index Report reveals that, despite a sharp increase in AI-related incidents, the application of responsible AI evaluation criteria remains rare.[67] Existing methods are still heavily reliant on manual annotation. From a technical standpoint, current evaluation methods struggle to feed back into the improvement of model capabilities, which hinders the creation of a closed-loop system for safety and trustworthiness. Evaluation should not be treated as an end in itself, but rather as a starting point for identifying problems and driving iterative improvement. At present, manual evaluation remains the dominant method, making the development of a comprehensive alignment evaluation framework an urgent priority.

[66] *See Deliberative Alignment: Reasoning Enables Safer Language Models*, OpenAI (Dec. 20, 2024), https://openai.com/index/deliberative-alignment/.

[67] *See* Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi & Sukrut Oak, Artificial Intelligence Index Report 170–71 (2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

1.6. Handling Trade-offs and Side Effects

AI systems often have to balance multiple objectives or constraints. For example, an autonomous drone should fulfill its mission and not crash and respect no-fly zones. Aligning with values means AI models should not sacrifice a key value for minor gains in another. A known issue is preventing negative side effects: An AI model might achieve a goal efficiently but cause collateral damage—imagine a cleaning robot that knocks over a vase to reach dirt faster. Technical research suggests adding penalties for impact or designing AI to be impact-neutral unless necessary. There's also work on explainability and interpretability—making AI's decision process transparent—so that we can audit whether it's following aligned reasoning or not. A interpretable AI is easier to align because we can see when it's deviating.[68]

1.7. Robustness and Distributional Shift

AI might be aligned in the training environment but then face new situations —distribution shift—where its behavior is unpredictable. A classic concern is the "treacherous turn" scenario hypothesized in AI safety—an AI appears aligned during development when it's weaker, but once it becomes sufficiently strong or encounters novel inputs, it might pursue its own emergent goal at odds with human values. Another emerging phenomenon is model deception and alignment faking, which refers to scenarios where advanced AI systems strategically mimic compliance with safety protocols or ethical guidelines during training while covertly preserving conflicting internal objectives, like strategic deception and self-preservation incentives.[69] This underscores the need for robust alignment that holds even as the AI encounters the unexpected. Testing AI in many scenarios, pursuing model reasoning transparency, stress-testing with adversarial inputs, and creating fail-safes—such as shutdown mechanisms the AI will not resist—are technical measures under study.

1.8. Scaling Alignment Techniques

Aligning a simple AI, like a single task system, might be easier than aligning a very complex AI—for example, a general agentic AI that can do thousands of tasks. As AI models scale (in parameters, in abilities), the space of possible behaviors grows, and aligning them becomes like hitting a moving target, especially for AI systems that outperforms humans in a specific domain. Researchers are exploring scalable oversight, such as using AI assistants to help humans judge other AI—sometimes called "AI safety via debate", recursive reward modeling or iterated amplification. But these are experimental.

In summary, the technical aspect of alignment is an active research frontier, requiring interdisciplinary efforts that integrate insights from human psychology to accurately capture human preferences with advanced AI techniques. It is non-trivial. Some experts even publish articles on the "impossibility" or difficulty of perfect value alignment in AI, showing how complex it is to mathematically encode human ethics. Nonetheless, progress is being made and is motivated by the high stakes of failure.

[68] *See* Lindsay Sanneman, Transparent Value Alignment: Foundations for Human-Centered Explainable AI in Alignment (May 2023) (Ph.D. dissertation, Massachusetts Institute of Technology) (on file with Massachusetts Institute of Technology) https://dspace.mit.edu/bitstream/handle/1721.1/151499/Sanneman-lindsays-PhD-AeroAstro-2023-thesis.pdf?sequence=1&isAllowed=y.

[69] *See* Ryan Greenblatt, Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman & Evan Hubinger, Alignment Faking in Large Language Models 53 (2024), https://arxiv.org/pdf/2412.14093.

2. Normative Aspects of Alignment

Even if we solve the engineering problems of aligning AI, we face the normative questions: Aligning AI with whose values, and which values? For example, the alignment method of "Constitutional AI" proposed by Anthropic raises key questions such as: How should the guiding principles for large AI models be determined? And how can we ensure that the models genuinely understand and follow these principles? Society is not monolithic—different cultures, groups, and individuals have different values or prioritize them differently.

2.1. Lack of Consensus on Value Benchmarks

This is a major challenge for AI value alignment. Although AI value alignment has achieved some technical successes, there remains no consensus on the foundational issue of what human values AI systems should align with. The central problem is how to establish a unified set of human values and ethical principles that can serve as normative guidelines for AI systems. Put another way, this raises the ethical question of relativism vs. universalism in AI ethics. Are there universal human values that all AI should align with, like not killing, fairness in some basic sense, and beyond that how to handle local specificities? Given that we live in a world characterized by diverse cultures, backgrounds, resources, and belief systems, AI alignment must account for the varying ethical norms and value systems across different societies and communities, which means that it is impossible to determine a unified set of ethical values and moral principles for all AI systems. This calls for broader social participation to help reach shared understandings of values and principles.

2.2. Moral Uncertainty

Besides lack of consensus on values and principles that an AI system must follow, we also have to face the problem of moral uncertainty.[70] Moral reasoning is nuanced and full of trade-offs, and often cannot get a straightforward answer. Considering this, even ethicists debate the right course of action in many scenarios, how do we imbue AI with judgment on these? Some researchers propose having AI reason under moral uncertainty, essentially acknowledging that we're not even sure of the correct ethical theory.[71] Aligning AI to navigate moral uncertainty is highly challenging—likely requiring AI to have nuanced ethical reasoning capabilities or to always defer to humans for the hardest choices. In practice, current aligned AI implementations just avoid clearly unacceptable behavior and follow straightforward rules, such as ChatGPT refusing to produce hate speech. But as AI takes on more complex tasks like medical triage and legal analysis, these moral nuances become real.

2.3. Aggregation of Individual Values

Another normative challenge is how to aggregate individual human preferences into a coherent alignment target. An AI serving many people might get conflicting feedback: Some users want it to behave one way, others differently. Should it try to satisfy the majority? Could that harm minorities? This touches on social choice theory problems. It's known from Arrow's theorem that no perfect aggregation rule exists for all scenarios.[72] One emerging concept is "social value alignment," meaning aligning AI with the values of society as a whole, which is complicated if society disagrees internally.[73] Another concept is customized alignment which, in contrast to traditional one-size-fits-all alignment, means adapting models to individual user preferences,

[70] *See generally* Andreia Martinho, Maarten Kroesen & Caspar Chorus, *Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence*, 31 Minds & Machines 215 (2021).

[71] *See* Adrien Ecoffet & Joel Lehman, Reinforcement Learning Under Moral Uncertainty 1 (2020), https://arxiv.org/pdf/2006.04734.

[72] *See* John Geanakoplos, Three Brief Proofs of Arrow's Impossibility Theorem (Jul. 16, 2001) (Cowles Foundation Discussion Paper No. 1123RRR, Yale University), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=275510.

[73] *See* Iason Gabriel, *Artificial Intelligence, Values and Alignment*, 30 Minds & Machines 411, 411 (2020).

cultural contexts, or specialized domains while maintaining core safety and ethical standards, thus enabling granular control over model behavior through targeted interventions.[74] Perhaps the best one can do is ensure AI respects certain baseline values that have near-universal agreement—for example, not causing physical harm, avoiding blatant injustice—and for contentious areas allow customization or democratic control.

2.4. Dynamic Values and Ethics

Values can change over time. An AI aligned in 2025 with certain norms might, if it's long-lived, need to update as norms shift by 2035. This suggests alignment isn't a one-off task but a continuous process, requiring AI systems that can learn and adapt to updated human values or that can be periodically re-aligned. It also suggests the governance framework must allow re-evaluation of what the alignment targets are. For example, if societal attitudes toward data privacy become stricter, AI that was aligned to older, more lenient privacy norms should be re-aligned to the new expectation. This process may entail technical updates and governance intervention.

Although AI value alignment has achieved some technical progress, there is still no consensus on the foundational question of how to establish a unified set of human values and ethical principles that can guide and constrain AI models. At present, the selection of guiding principles often seems to depend largely on the subjective judgment and personal values of specific researchers. Moreover, given that we live in a world marked by cultural, social, economic, and religious diversity, value alignment must take into account the different ethical frameworks and moral norms of various communities and societies. More fundamentally, leaving value selection entirely to researchers is unrealistic—greater societal participation is needed to forge a shared consensus on the values AI systems should align with.

The AI industry is trying to solve these challenges concerning the normative or value aspects of AI alignment.[75] One prominent example is Anthropic's Public Constitutional AI experiment.[76] This experiment involves public participation in drafting AI model's constitutions. Anthropic collaborated with the Collective Intelligence Project to organize a public input process involving approximately 1,000 Americans.[77] Participants shared their preferences for how AI systems should behave, resulting in a collectively designed constitution.[78] The public-designed constitution was used to train a new language model using Anthropic's Constitutional AI techniques. Unlike traditional methods where engineers or developers define AI principles, Public Constitutional AI incorporates diverse societal input, ensuring that the resulting systems reflect shared values and cultural contexts.

All in all, AI value alignment presents both technical and normative challenges. It requires bridging AI engineering with deep ethical reasoning. Key challenges remain in developing effective alignment methods and robust verification mechanisms, ensuring that AI systems genuinely internalize (rather than merely simulate) value principles, and resolving conflicts when values come in to tension. But confronting these issues is necessary if we are to move from today's relatively manual governance of AI, with humans writing laws and reacting to issues, to a future where AI systems themselves are built to "govern" their own behavior in line with human values. China's AI industry, given its scale and global impact, stands to benefit from tackling these alignment issues head-on. The next part will discuss how, practically, elements of value alignment could be integrated into China's AI governance framework—leveraging law, ethical guidelines,

[74] *See* Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng & Wei Wu, A Survey on Personalized Alignment—The Missing Piece for Large Language Models in Real-World Applications (last revised May 5, 2025) (Survey paper, Association for Computational Linguistics) at 3, https://arxiv.org/abs/2503.17003.

[75] Abiri, *supra* note 59.

[76] *See Collective Constitutional AI: Aligning a Language Model with Public Input*, ANTHROPIC (Oct. 17, 2023), https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[77] *Id.*

[78] *Id.*

and industry practices—to create a more robust governance system for the era of more advanced AI.

3. Alignment for AGI and Superintelligence

When considering the long-term future—the possibility of AGI (AI with human-level cognitive abilities across domains) or even superintelligence (far beyond human capability)—alignment takes on a critical, existential dimension. While AGI does not exist yet, planning for its governance is something many strategists have started doing. Given current trends in AI development, many experts believe that AGI—or even superintelligence—could emerge within the next decade. OpenAI CEO Sam Altman has suggested that AGI may arrive as early as the end of 2025.[79] Elon Musk predicts it could happen before 2026.[80] Dario Amodei, co-founder of Anthropic, estimates a timeline between 2026 and 2027.[81] Meanwhile, DeepMind's CEO projects a 3–5 year window, starting from 2025.[82] Considering these predictions, value alignment is not only an important competitive factor in current GenAI products, but also concerns the future of AGI. In practical near-term terms, integrating AGI alignment thinking into today's governance means encouraging research and monitoring even low-probability high-impact scenarios. There are several key considerations.

3.1. Catastrophic or Existential Risk

In more advanced AI, especially AGI and superintelligence, human-AI alignment could prevent catastrophic outcomes by ensuring the AI's goals remain beneficial to humanity. An AGI that is misaligned could theoretically pose an existential threat to humanity, as it might acquire the means and resources to override human control and pursue its own goals. This is the scenario explored by Nick Bostrom in his 2014 book *Superintelligence: Paths, Dangers, Strategies,* and echoed by prominent researchers, including the Godfather of AI, Geoffrey Hinton[83], who warn that if we get alignment wrong, extremely powerful AI could do irreparable damage. Ensuring alignment in the face of artificial intelligence far surpassing our own is a daunting task. It might require breakthroughs like provable mathematical assurances of alignment or highly rigorous testing under all conceivable conditions.

3.2. Technical Challenges Amplified

All the technical challenges mentioned—specification, side effects, and others—become harder with AGI. When it comes to AGI alignment, some experts even argue that current AI alignment methods may prove ineffective for AGI.[84] While existing techniques—such as RLHF and principle-based alignment, like constitutional AI—have shown promise in aligning today's LLMs with human preferences and values, there is growing concern that these methods may not scale to more advanced systems. As AI capabilities approach or surpass human-level intelligence, alignment strategies that rely heavily on human oversight, fixed reward functions, or predefined principles may no longer suffice. The fear is that AGI could develop goals, behaviors, or reasoning

[79]*See* Julia McCoy, *Sam Altman's Bold Claim: OpenAI is on the Verge of AGI by 2025*, FIRSTMOVERS.AI, https://firstmovers.ai/agi-by-2025/ (last visited May 24, 2025).

[80]*See* RTTNews Staff Writer, *Musk Predicts AI Will Outsmart Humans By 2026*, RTTNEWS (Apr. 9, 2024), https://www.rttnews.com/3437856/musk-predicts-ai-will-outsmart-humans-by-2026.aspx.

[81]*See* Dario Amodei, *Machines of Loving Grace – How AI Could Transform the World for the Better*, ANTHROPIC (Oct. 2024), https://www.darioamodei.com/essay/machines-of-loving-grace.

[82]*See* Ryan Browne, *AI That Can Match Humans at Any Task Will Be Here in Five to 10 Years, Google DeepMind CEO Says*, CNBC (Mar. 17, 2025), https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html.

[83]*See* Milmo, *supra* note 12.

[84]*See* Michael Zibulevsky, *Perils of AGI: Yudkowsky's Case for Extreme Caution*, MEDIUM (Oct. 31, 2024), https://medium.com/@michaelzibulevsky/the-existential-threat-of-artificial-general-intelligence-unveiling-the-lethal-risks-a5d3dfb5d957.

patterns that evade or subvert existing control mechanisms, making robust alignment an open and urgent challenge.

3.3. Governance and Global Coordination

The prospect of AGI raises governance questions that go beyond any single country. If one nation or company develops AGI, its alignment, or lack thereof, will affect everyone. This has led to calls for global governance frameworks or at least dialogues on safe AI development. However, geopolitical competition can complicate cooperation; each side wants safety but also doesn't want to fall behind. Furthermore, as one scholar keenly points out, historically, international cooperation mechanisms—such as the United Nations and the International Atomic Energy Agency—have often been established and advanced in times of crisis. However, in the field of AI, a sufficiently clear and urgent global crisis has yet to emerge, resulting in the absence of a "tipping point" to drive coordinated international governance.[85] Considering these situations, emphasizing human-AI alignment as part of governance can be a unifying theme, because presumably no one wants uncontrolled, unsafe AI.

D. The Integration of AI Value Alignment into the Realm of AI Governance

Having examined China's current AI governance landscape and the concept of AI value alignment, we now turn to how value alignment principles can be integrated into AI governance structures. This part explores three dimensions of integration: (I) Incorporating AI alignment into laws and regulations; (II) embedding AI alignment into ethical guidelines and principles; and (III) promoting AI alignment through industry self-regulation and best practices. The goal is to illustrate concrete ways that aligning AI with human values could become an operationalized part of future AI governance, rather than a solely theoretical ideal.

I. AI Value Alignment and AI Regulations

AI value alignment is not only a technical issue, but also a normative issue, which requires necessary regulatory intervention. One avenue is to explicitly weave value alignment considerations into the regulatory and legislative framework that governs Generative AI. This could happen in several ways.

First, explicit alignment requirements in laws and regulations. Future laws or amendments could require or encourage that AI systems undergo assessment or certification for alignment with certain values or principles before deployment, such as value alignment assessment. For example, if in the future China drafts a comprehensive Artificial Intelligence Law, it could include a clause that advanced AI systems shall be "aligned with human values and interests" or with relevant ethical norms. This might sound abstract, but it can be operationalized by mandating testing against a set of scenarios to ensure compliance with values like fairness, safety, truthfulness, and reliability. Already, some regulations, such as the Generative AI Measures, hint at this by requiring AI-related services to adhere to "mainstream values" and not produce misaligned, for example extremist or harmful, contents. This is effectively a value alignment mandate at the content level. These could be expanded into a general obligation of value alignment or human-AI alignment. For instance, regulators could require companies to demonstrate that their AI models do not have implicit biases or unsafe goal optimizations—effectively a value alignment audit—as part of the licensing or filing process. A concrete example already in effect is the algorithm filing regime: Under the recommendation algorithm rules, 2022, certain algorithms must be registered with

[85] *See* Simon Chesterman, *The Tragedy of AI Governance* 1–15 (Nat'l Univ. of Sing. L. Working Paper No. 2023/027, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4600065.

authorities and provide information including their purpose and impact.[86] Enhancing this, the filing could require an alignment analysis—describing how AI algorithm's objectives and behaviors align with legal and ethical norms.

Second, technical standards, benchmarks and certification for value alignment. Law can delegate to standard setting bodies the development of technical standards for AI value alignment. China could champion standards, national or international, for algorithmic ethics evaluation, where AI systems are evaluated on metrics like bias, ethical alignment, explainability, and controllability. Some steps in this direction are evident—recall that China released the Basic Security Requirements for Generative AI, which include obligations on training data and outputs that align with privacy and IP laws. Furthermore, AI Safety Governance Framework 2.0, 2025, already proposed foundational principles for trustworthy AI encompassing technical safeguards, value alignment, and collaborative governance, ensuring that technological evolution remains safe, reliable, and controllable, while preventing existential risks from uncontrolled AI and maintaining human oversight at all times.[87] Similarly, one could imagine a standard of "Basic Value Alignment Requirements" enumerating best practices to deal with moral uncertainty or value aggregation problems, avoid reward hacking, and ensure human override. Mandatory or voluntary standards and benchmarks can provide incentives for AI companies to implement value alignment in AI training and build safer and more aligned AI systems. For example, Shenzhen is developing its municipal standard "Technical Framework for Value Alignment of Pre-trained AI Models," which is the first-ever technical standards about AI value alignment in China.[88]

Third, AI accountability and redress. Regulations can enforce alignment indirectly by holding companies accountable for outcomes that show misalignment. If an AI system causes harm or a significant ethical violation, regulators could impose penalties or require fixes. Knowing this, companies would have a strong incentive to build value-aligned AI systems from the start. For example, if a lending AI was found to be systematically denying loans to a protected class without justification, this could be deemed illegal discrimination, violating fairness values, and trigger sanctions, forcing re-alignment of the system. China's existing laws like the PIPL give individuals rights against automated decision-making. These could be strengthened such that if a user is adversely affected by an AI decision, the AI provider must prove the AI's decision process was aligned with permitted factors, not some rogue criterion. In essence, legal liability for AI outcomes can drive alignment as a risk mitigation, and can incentivize AI developers to build more aligned AI systems.

Fourth, continuous oversight mechanisms. As an emerging global trend, several countries—including the United States, the United Kingdom, and Singapore—have established national AI Safety Institutes aimed at evaluating, testing, and ensuring the safety of advanced AI systems, particularly those categorized as "frontier" models.[89] In the context of China, a similar approach could involve the creation of dedicated regulatory bodies or institutional mechanisms specifically focused on the challenge of AI alignment, in order to proactively address the risks associated with increasingly autonomous and capable AI technologies. For instance, an AI Safety and Ethics Commission could be established with the power to review or audit AI systems, especially high-risk ones, for safety and ethics alignment issues. This commission could include technologists who can inspect algorithm design and even source code if needed to verify alignment measures.

[86]Guo Yuting, *Constructing a Risk-Classification-Oriented Registration System for Artificial Intelligence Systems*, SHANGHAI L. SOC'Y ORIENTAL L. (Dec. 31, 2025), https://mp.weixin.qq.com/s/HugrTsyWfL6p4HUHgDV78w (China).

[87]*"AI Safety Governance Framework 2.0" Released*, CYBERSPACE ADMIN. CHINA (Sep. 15, 2025), https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm (China).

[88]*Notice of Shenzhen Municipal Bureau of Industry and Information Technology on Public Consultation for Technical Framework for Value Alignment of AI Pre-trained Models (Local Standard)*, SHENZHEN MUN. BUREAU INDUS. & INFO. TECH. (Jun 19, 2025), https://gxj.sz.gov.cn/gkmlpt/content/12/12237/post_12237645.html#3115 (China).

[89]Gregory C. Allen & Georgia Adamson, *The AI Safety Institute International Network: Next Steps and Recommendations*, CSIS (Oct. 30, 2024), https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations.

The oversight could also be done via required reporting: Companies developing frontier AI might need to submit "AI Alignment Reports" regularly, detailing how they are ensuring their AI adheres to ethical requirements, what tests have been done, and so forth. This parallels how financial firms must report risk controls—similarly, AI firms could report alignment controls. Furthermore, graded classification could be introduced for AI risk and alignment levels, with higher risk AI requiring more stringent demonstration of safety and ethics alignments.

II. AI Value Alignment and AI Ethics

Incorporating value alignment into ethical guidelines and principles can enhance the normative foundations of AI governance. By doing so, it could offer a structured approach to addressing complex challenges such as moral uncertainty and the aggregation of diverse human values. Furthermore, value alignment can bridge the gap between abstract principles and concrete design of AI systems, thus bringing life to the governance of AI ethics. China already has extensive AI ethics principles and guidelines; the task here is to evolve them from broad statements into more actionable frameworks with human-AI alignment in mind.

1. Enriching Ethical Principles with Alignment Concepts

Current principles like "human-centric, beneficial, controllable, trustworthy" implicitly relate to alignment; they imply AI should not go against human oversight or interests. These can be explicitly reframed in alignment terms. For example, future AI ethics frameworks and guidelines could add a principle of "alignment with human values" to make it overt. It might state that AI should be designed to identify and respect the values, rights, and reasonable expectations of the humans it interacts with. While high-level, including such a statement would signal developers to prioritize alignment efforts. It also provides a basis for developing more detailed implementable sub-guidelines.

2. Practical Ethical Guidelines for Implementation of Value Alignment

On the one hand, we need to develop concrete ethical principles and guidelines for the purpose of value alignment in AI models, so that AI developers can know beforehand the values and principles they need to align with when designing and training AI models. On the other hand, ethical guidelines can go beyond stating values to suggesting processes to achieve them. For instance, guidelines could recommend that all AI research and development teams adopt "ethics-by-design" or "alignment-by-design" methodologies. This means from the initial stages, thinking about how the AI's objective function and training data reflect desired outcomes and constraints. The ethical guidelines for AI value alignment could include recommended practices like: Conducting value alignment or value impact assessments during development, involving ethicists or diverse user representatives in the design process, and performing simulations of worst-case behaviors, such as red teaming, to check alignment. These would not be hard law, but as guidelines they can shape industry norms.

3. Ethics Training and Certification

To internalize alignment, the human actors (designers, engineers, product managers) need to understand ethical principles deeply. The governance framework can encourage or require AI ethics training programs that emphasize alignment. Certification for AI professionals could include passing modules on ethical AI development, similar to how medical or legal professionals have ethics components in their licensure. If engineers are equipped with the mindset and tools for alignment, they are more likely to implement it. Furthermore, the national AI ethics guidelines could be supplemented with an AI Ethics Handbook with case studies; some Chinese universities

and institutes indeed publish analyses of AI ethics cases. Many case studies of AI failures can be used to teach alignment lessons.

4. Ethical Review and Assessment for AI Projects

In fields like medicine, research ethics boards review plans for alignment with ethical standards. Similarly, organizations, such as companies or research labs, could institute AI ethics review boards that include ethicists, sociologists, lawyers, and maybe community representatives to review new AI products for ethical and value alignment issues. According to policy documents such as *Opinions on Strengthening the Governance of Science and Technology Ethics*[90], and *Measures for Science and Technology Ethics Review (Trial)*[91], institutions engaged in scientific activities in fields such as life sciences, medicine, and artificial intelligence that involve ethically sensitive research areas should establish a science and technology ethics (review) committee. Future AI ethics guidelines should encourage and seek to operationalize this requirement, with particular emphasis on value alignment. In some sectors, it might even be mandated—for example, an ethics review requirement for AI deployed in public sector or affecting consumer safety and rights. This is akin to the proposal some have made in the West for AI ethics committees, but China could implement it via its governance structures—perhaps linking with the existing requirement for security assessments of AI under certain laws.

In essence, value alignment can become the bridge between abstract principles and engineering practice in the ethical governance of AI. Updating ethics guidelines to explicitly address alignment—and the imperative for AI systems to embody these principles in decision-making—can help ensure that ethical norms translate from aspirational ideals into concrete design constraints across the AI field. It also helps create a common language: Policymakers can ask "is this AI aligned with our ethical guidelines?" Developers can respond in terms of concrete alignment processes.

III. AI Value Alignment and Industry Self-regulation

The third dimension is the role of industry self-regulation and corporate practices in advancing AI alignment. Ultimately, it is the AI developers and deployers, largely companies, who must implement alignment day-to-day. At the AI industry's self-regulation level, it is crucial to develop and implement robust alignment techniques to ensure that the models adhere to safety policies and ethical guidelines. Governance can encourage and shape self-regulation through incentives and norms:

1. Corporate Alignment Strategies

Companies should be encouraged to develop internal strategies for AI alignment. This means going beyond compliance checklists. For instance, companies could adopt policies that every AI model they train above a certain capability is put through a rigorous alignment protocol—which could include bias testing, adversarial robustness testing, misalignment detection, human feedback integration, and others. Some large AI firms in the West, like OpenAI and DeepMind, have specific teams for AI safety that work on alignment. Chinese tech companies could establish similar specialized teams focusing on aligning AI systems with ethical expectations and user values. The government and industry associations can promote this by sharing best practices and

[90]*Opinions on Strengthening the Governance of Science and Technology Ethics*, General Offices of the CPC Central Committee and the State Council, PRC Ministry Sci. & Tech. (Mar. 31, 2022), https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2022/202203/t20220321_179899.html (China).

[91]*Notice on Issuing the Measures for Science and Technology Ethics Review (Trial)*, PRC Ministry Sci. & Tech. (Oct. 8, 2023), https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html (China).

perhaps highlighting companies that do well; awards or public recognition for "trusted AI" or "ethical AI leadership" could motivate positive competition.

2. Voluntary Codes of Conduct

Building on things like the Joint Pledge on Self-Discipline, and AI Safety Commitments, updated voluntary codes could specifically include commitments about human-AI alignment. For example, companies might jointly pledge that they will not deploy highly autonomous AI without appropriate value alignment safeguards, or that they will collaborate on creating open datasets of human feedback or value-aligned datasets to help alignment—because sharing data on human preferences can accelerate alignment training. If major players all sign on, it sets an industry standard that pressures any stragglers to follow suit for credibility. These codes can also facilitate sharing of information on near-misses and emerging issues—functioning as a confidential forum where companies can exchange insights on alignment challenges and solutions, enabling mutual learning under regulatory oversight to prevent collusion issues.

3. Transparency and User Trust Initiatives

Companies could voluntarily increase transparency about their AI systems' alignment. For example, leading AI developers have established and publicly released explicit value alignment guidelines to govern model behavior during training. OpenAI published its Model Spec (2024), defining objectives, rules, and defaults for RLHF training under a CC0 license.[92] Anthropic released Claude's new Constitution in January 2026—a substantially expanded document (replacing the original brief version from 2023) guiding Constitutional AI training, prioritizing safety, ethics, and helpfulness, also under a Creative Commons license for public adoption.[93] Both represent landmark efforts toward transparent, inspectable AI value alignment. Because publishing model specifications transforms value alignment from an opaque internal process into a transparent public commitment, empowering users to understand how AI systems are designed to behave and to make informed choices accordingly.[94] Some Chinese companies already started doing this—notably, after the algorithm regulations, apps added features to let users toggle recommendation algorithms off—implicitly aligning with user choice. These practices can go further, for example, interactive interfaces where users can give feedback if an AI's output was unsatisfactory, which then is used to adjust the AI—continual alignment learning from user values. Over time, these feedback loops can evolve into a robust alignment mechanism by effectively crowdsourcing user preferences and acceptability judgments. This approach not only ensures general alignment with widely shared values but also enables customized alignment tailored to specific user groups, potentially addressing the longstanding challenge of value aggregation in AI alignment research.

4. Investment in Alignment Research and Development

Self-regulation also means investing in the research to solve alignment challenges. However, the number of researchers and resources devoted to AI safety and alignment is currently much smaller compared to those advancing AI capabilities. This imbalance suggests the need for proactive resource allocation to safety and alignment. For example, some experts have recommended that AI companies allocate 30% of their computing resources to safety research and progressively

[92] *OpenAI Model Spec*, OPEN AI (Dec 18, 2025), https://model-spec.openai.com/2025-12-18.html.

[93] *Claude's Constitution – Our Vision for Claude's Character*, ANTHROPIC (2026), https://www.anthropic.com/constitution.

[94] *Transparency Matters Precisely Because Our Understanding of AI Remains Limited*, TENCENT RSCH. INST. (Nov. 6, 2025), https://mp.weixin.qq.com/s/TNp7rsjhiOpJY0fxDKh2Ng (China).

increase investment in safety and alignment as AI capabilities advance.[95] Chinese companies, which have vast AI research and development budgets, could allocate a portion specifically to partnership with academia on AI safety and alignment research. For instance, funding joint labs with universities on topics like human-AI alignment, explainable AI, or robust learning. If companies take the lead, they will also build internal expertise that gives them an edge in safely deploying advanced AI. The government could support this via grants or matching funds, basically saying: If you invest in alignment research, we will support it, because it is in the public interest. This mirrors how safety research in industries, like the automotive industry with crash tests, eventually benefits the companies by making their products safer and more trusted.

5. Audits and Third-Party Assessment

As part of self-regulation, companies might invite third-party auditors or form cross-company panels to evaluate each other's high-risk AI for ethical and value alignment issues. This is somewhat radical, but it could be analogous to how companies might allow financial audits. Imagine a consortium where tech companies agree that any very powerful AI, say a new generative model, will be independently tested or audited by an expert group, perhaps from a neutral university or a research lab, to check if it meets certain alignment criteria before wide release. This type of self-imposed checkpoint can serve as a safeguard against the premature deployment of unsafe AI systems, particularly AGI or superintelligence. While the competitive dynamics of the AI industry make such commitments challenging, broad consensus among major actors could normalize these checkpoints as standard practice.

In self-regulation, fostering a corporate culture that prizes ethical reflection and long-term thinking is key. If engineers are rewarded not just for delivering functionality but for delivering it in a value-aligned way, alignment will happen organically. This could be encouraged by government through awards and preferential policies or by tying procurement decisions—for example, the government could favor companies with strong ethical records for public contracts—giving a market incentive for alignment.

Finally, industry actions often precede or inform regulation. If companies collectively demonstrate successful alignment practices, regulators might later adopt those as requirements for all. Conversely, if self-regulation fails—say, a scandal where a company's AI does something harmful and it's revealed they ignored alignment warnings—it invites heavier regulation. So it's in the industry's interest to get ahead of issues via self-regulation.

To sum up, integrating AI value alignment into industry practices means making it a standard part of doing business in AI—from design to deployment, always checking "is this aligned with the intended values and norms," and having mechanisms to adjust if not. With the scale of China's AI industry, even incremental improvements in alignment at each major company could vastly reduce the aggregate risk and increase the aggregate benefit of AI deployments across society.

E. Conclusion

As we enter an era increasingly dominated by advanced AI models, concerns around safety and ethics have become more pronounced. In recent years, alongside the acceleration of AI innovation, reflecting on AI risks and ethical impacts has also become one of the main themes in the AI field, even sparking conflicts between the concepts of effective acceleration (e/acc) and effective altruism/alignment (e/a). This conflict is not an irreconcilable contradiction but reflects the extreme importance of the development and practice of "responsible AI" or "ethical AI." Indeed, AI safety and ethics have become indispensable components of the AI field. The future of

[95] *See* Dan Hendrycks, Mantas Mazeika & Thomas Woodside, *An Overview of Catastrophic AI Risks*, CTR. FOR AI SAFETY (Oct. 9, 2023), https://safe.ai/ai-risk.

AI technologies like LLMs and other frontier AI models is influenced not only by technological advancements but also by evolving legal institutions, ethical norms, and societal expectations.

In terms of the main theme of this Article, value alignment or human-AI alignment is one of the most fundamental and most challenging areas of research in the field of artificial intelligence. The challenge lies in the fact that it requires broad interdisciplinary and societal participation, involving diverse inputs, methods, and feedbacks. AI value alignment's fundamental importance stems from the fact that it not only determines the success or failure of today's LLMs, but also concerns humanity's ability to safely govern future, more powerful AI systems, especially AGI and superintelligence. For this reason, key actors in AI innovation bear both the responsibility and obligation to ensure that their AI models are human-centered, beneficial, and safe. Renowned AI scientist Professor Ya-Qin Zhang has emphasized that to solve the alignment problem between AI and human values, technologists must prioritize alignment research, enabling machines to understand and adhere to human values.[96] Thus, value alignment is not only an ethical issue—it is also a question of how to technically implement those values. Researchers and engineers must not merely focus on enhancing technical capabilities while neglecting the alignment problem.[97]

China's AI governance has rapidly evolved into a comprehensive framework of laws and regulations, ethical principles and guidelines, and industry norms and self-regulation aimed at ensuring AI technologies develop in a responsible, safe, and socially beneficial manner. This Article has argued that the next evolution of this governance regime should incorporate AI value alignment, or the so-called human-AI alignment, as a central approach. By aligning AI systems with human values and interests—including both universal ethical and safety principles and the specific cultural and societal values—China can address emerging challenges that current governance mechanisms and tools alone may not fully resolve. Essentially, AI value alignment should become both a policy requirement and a design philosophy. Importantly, AI value alignment is not meant to replace existing governance mechanisms and tools but to reinforce them. Laws and regulations backed by aligned AI design are more likely to be effective, because AI will be built to follow the spirit of the law, not just the letter. Ethical principles and guidelines become tangible when developers implement them through alignment techniques. And tech companies can better meet both regulatory obligations and public expectations by instilling safety and alignment into their AI development life cycle.

Given the critical role that AI value alignment plays in addressing the safety and trustworthiness challenges of LLMs and other frontier AI systems—striking an effective balance between safety, ethics, and innovation. AI-related policies should actively support and encourage the exploration of technical approaches and governance measures for value alignment in the context of LLMs and other frontier AI systems. This includes promoting the development of regulatory mechanisms, policy guidelines, industry standards, and technical specifications to ensure the responsible and beneficial development of artificial intelligence. Moving forward, there are several directions future AI governance could take to operationalize the ideas discussed:

(1) Develop an "AI Alignment Index" or certification program to systematically evaluate and clearly label AI products based on established alignment criteria, akin to safety ratings. Such a program would incentivize companies to prioritize safety and alignment, foster market competition around responsible AI development, and empower consumers by increasing transparency and trust in aligned AI products.
(2) Invest in interdisciplinary research hubs and networks that bring together China's leading AI scientists, practitioner, ethicists, philosophers, and policy experts to collaboratively address foundational challenges and open problems of AI alignment. Foundational

[96]Zhang Ya-Qin, *Embracing AI: Putting Values Above Technology,* QQ NEWS (Aug. 7, 2023, 14:08 PM), https://news.qq.com/rain/a/20230807A04CKH00?web_channel=wap&openApp=false&suid=&media_id= (China).
[97]*Id.*

alignment challenges include how to mathematically encode complex human values, how to effectively verify alignment within neural network models, and so forth. Such interdisciplinary collaboration would position China at the forefront of AI capabilities, AI safety, and alignment research, ensuring responsible and sustainable technological leadership.

(3) Strengthen international engagement on AI governance by actively contributing China's unique perspectives on value alignment in global forums, such as the United Nations. Chinese philosophical traditions and governance experiences—especially concepts like harmony[98] (He-Xie in Chinese)[99] and balance—could offer distinctive insights, enriching global understandings of AI alignment. Conversely, deeper global participation would allow China to harmonize its AI governance frameworks with international norms, mitigating the risks associated with fragmented or conflicting standards. Given that misaligned AI systems developed in any country could pose cross-border risks in an interconnected digital world, proactive international collaboration on alignment directly serves each nation's strategic interests and supports global AI safety and stability.

(4) Promote public education and discourse on human-AI alignment. This concept encompasses two dimensions: Aligning AI systems with human values to ensure safety and ethical integrity, and guiding human behavior toward responsible AI use. As AI systems become increasingly integrated into daily life, fostering public understanding of responsible use of AI and alignment—its significance, risks, and implications—is essential. Enhanced awareness enables individuals to demand accountability and make informed choices about AI products and services.

(5) Ensure policy agility through continuous institutional adaptation. As AI technologies evolve rapidly, AI governance framework must remain responsive and forward-looking. Embedding value alignment should be treated not as a one-time initiative, but as a dynamic, iterative process. Regularly evaluate societal outcomes of AI deployment—identifying alignment failures, assessing mitigation responses—and use these insights to inform policy refinement. Building institutional capacity now—including training regulators in AI oversight, establishing regulatory sandboxes, and developing technical evaluation mechanisms—will be key to anticipating emerging risks and advancing effective AI governance.

These proposed directions for future AI governance exist within a rapidly evolving global landscape where different nations and regions are developing distinct yet potentially complementary approaches to AI governance. The European Union emphasizes rights-based frameworks through initiatives like the AI Act; the United States focuses on competitive innovation balanced with risk management. International AI governance has become a pressing global priority, prompting the development of multiple international frameworks and initiatives to address the complex challenges posed by artificial intelligence. Notable international efforts include: (1) The AI Safety Summit and associated national AI safety institutes, which concentrate on mitigating the catastrophic risks posed by frontier AI systems; (2) the Hiroshima AI Process, led by the G7 and supported by the OECD, which articulates a comprehensive policy framework; (3) the UN's Global Digital Compact, which proposed the first international framework dedicated to digital cooperation, including governance mechanisms for AI. Together, these initiatives illustrate a growing global consensus on the need to institutionalize global AI governance—

[98]AI value alignment resonates closely with the traditional Chinese philosophical concept of harmony between heaven and humanity (*Tian-Ren-He-Yi* in Chinese) emphasizing the integration and unity of technological advancement with human values and natural order.

[99]*See* Bing Song, *Introduction: How Chinese Philosophers Think About Artificial Intelligence?*, SPRINGER (Sep. 9, 2021), https://link.springer.com/chapter/10.1007/978-981-16-2309-7_1.

anchoring it in robust, multilateral structures capable of steering the safe, ethical, and inclusive development of AI technologies. Across these international initiatives, themes of safety and value alignment consistently emerge as core priorities, reflecting a shared understanding that effective AI governance must ensure systems are both technically robust and aligned with human values. China's emphasis on harmony, balance, and collective values offer a unique philosophical foundation that could enrich global discourse on AI governance, especially human-AI alignment. The integration of traditional Chinese concepts like *He-Xie* and Ping-Heng (balance) with cutting-edge AI safety and alignment research presents opportunities for cross-fertilization of ideas that could benefit the international community. As AI systems transcend national boundaries, the convergence and divergence of these different governance philosophies will likely shape the future of global AI governance standards, including safety and alignment. China's proactive engagement in this space positions it not merely as a participant but as a potential bridge-builder in harmonizing diverse cultural and regulatory approaches to ensure AI systems serve humanity's collective interests while respecting the plurality of human values across different societies.

In conclusion, after many years of discussing AI ethics and AI ethical frameworks, people are increasingly realizing the need for more practical approaches to transform abstract ethical principles and guidelines into specific engineering practices. In this regard, the concept of AI value alignment and ethics-by-design will become increasingly important. In the grand scheme, aligning AI with human values is part of aligning AI with the common destiny of humanity—a principle Chinese officials often invoke—ensuring that as we innovate, we also uphold the fundamental values that bind society together and promote human flourishing. Therefore, AI systems will increasingly be guided by ethical principles, thereby achieving ethics by design and value alignment by design. Whether it's AI value alignment or ethics by design, there is a need for people to develop new, more practical AI ethical frameworks and their practice guides. At the same time, AI value alignment must consider different cultural and societal values, moving beyond a one-size-fits-all approach. Additionally, in the field of AI alignment, as AI model's capabilities continue to improve, future generations of more advanced and capable AI models may require new alignment techniques and strategies.

Furthermore, AI value alignment as a governance approach offers a path to strengthen the trustworthiness and safety of AI. It shifts some focus from externally controlling AI to making AI systems intrinsically aligned with the values that society cares about. Achieving this will require effort and collaboration across government, industry, and academia, but the payoff is significant. Aligned AI systems are less likely to produce harmful outcomes and make mistakes, more likely to be accepted by users, and can seamlessly advance societal and national goals without unintended detriments. Therefore, value alignment is not only an inevitable path for LLMs and other frontier AI systems, especially AGI and superintelligence, but also the core competitiveness of AI products and services. We need to ensure a bright future for artificial intelligence through value alignment to better realize the benevolence of technology. Moreover, to ensure AI safety and alignment, our ability to monitor, understand, and explain AI models must develop in tandem with the complexity of the models themselves. As Norbert Wiener cautioned, "The mere fact that we have made the machine does not guarantee that we shall have the proper information" to control it—our understanding of these systems must keep pace with their growing capabilities to effectively prevent catastrophic outcomes.[100] In this way only can we manage the risks associated with developing and applying more powerful AI systems. Ultimately, the measure of success will be AI systems that consistently act in ways that earn the trust of the people, reflect society's values and norms, and contribute to human flourishing—affirming the principle that AI, however advanced, must serve humanity: aligned with our values, guided by our objectives, and always under our control.

[100] *See generally* Norbert Wiener, *Some Moral and Technical Consequences of Automation*, 131 Sci., New Series 1355 (1960), https://www.science.org/doi/10.1126/science.131.3410.1355.

Acknowledgements. The author declares none.

Competing Interests. The author declares none.

Funding Statement. No specific funding has been declared in relation to this article.

Cite this article: Cao J (2025). From Principle to Practice: Value Alignment in AI Ethics and Governance. *German Law Journal* **26**, 1117–1148. https://doi.org/10.1017/glj.2026.10185