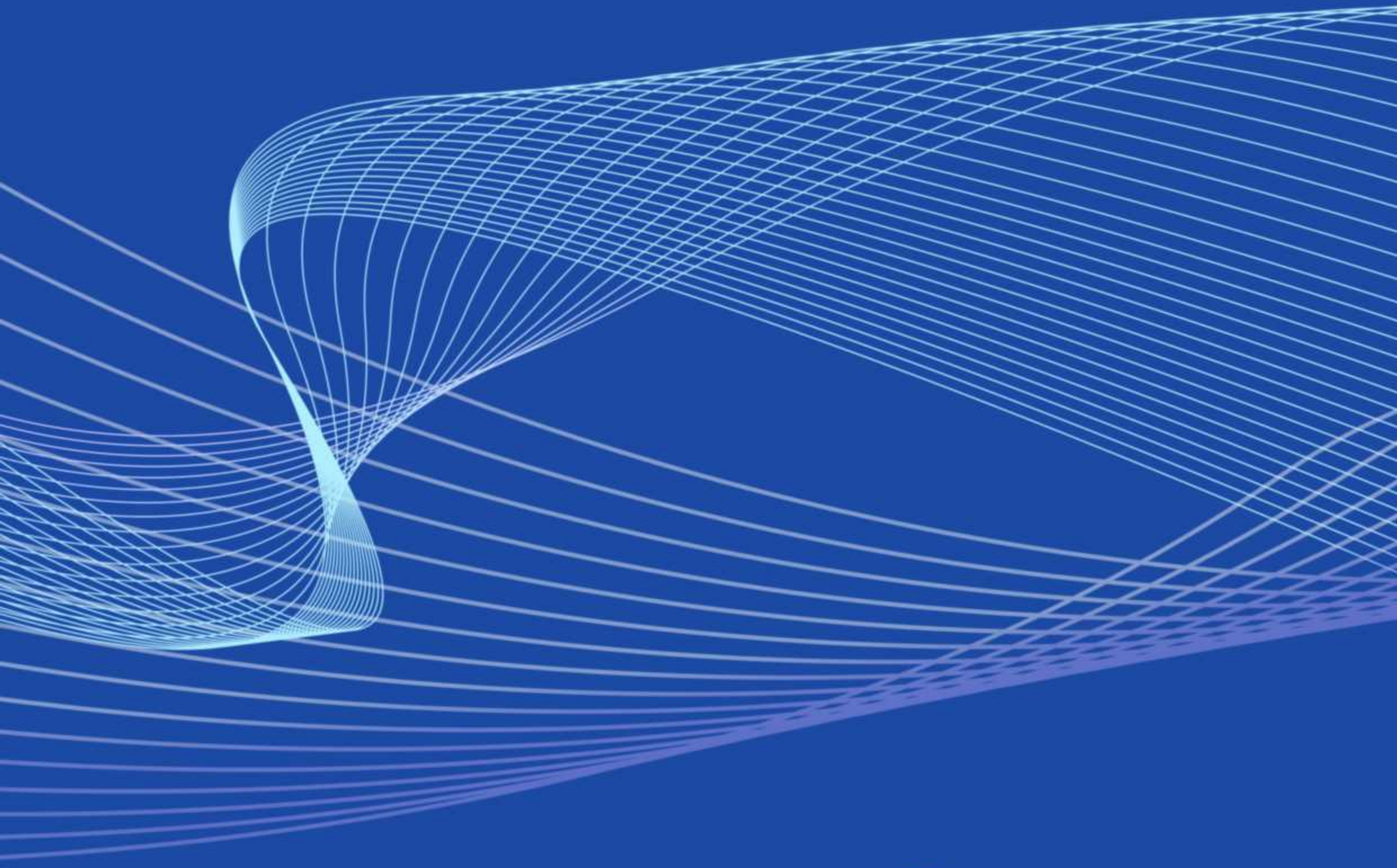
State of AI Safety in China

October 2023



CONCORDIA AI
安远AI

State of AI Safety in China

October 2023

Please direct inquiries to: info@concordia-ai.com

Executive Summary

Amid the rapid evolution of the global artificial intelligence (AI) industry, China has emerged as a pivotal player.[1] From advancing regulations on generative AI and calling for AI cooperation at the United Nations (UN), to pursuing technical research on AI safety and more, China's actions on AI have global implications. **However, international understanding of China's thoughts and actions on AI safety remains limited.** This report aims to close that knowledge gap by analyzing China's domestic AI governance, international AI governance, technical AI safety research, expert views on AI risks, lab self-governance methods, and public opinion on AI risks.

China has developed powerful domestic governance tools that, while currently not used to mitigate frontier AI risks, could be employed that way in the future.[2] Existing Chinese regulations have created an algorithm registry and safety/security reviews for certain AI functions, which could be adapted to more directly deal with frontier risks. Notably, an expert draft of China's national AI law attempts to regulate certain AI scenarios by building upon the algorithm registry to create licenses for more risky cases, among other policy tools.[3] The science and technology (S&T) ethics review system requires ethics reviews during the research and development (R&D) process for certain AI use-cases, though the system is still under construction and implementation details are yet to be clarified. While current domestic standards on AI safety are mostly oriented towards security and robustness concerns, China's top AI standards body referenced alignment in a 2023 document, suggesting growing attention towards frontier capabilities.

In the international arena, China has recently intensified its efforts to position AI as a domain for international cooperation. In October 2023, President Xi Jinping announced the new Global AI Governance Initiative (全球人工智能治理倡议) at the Third Belt and Road Forum for International Cooperation, setting out China's core positions on

[1] For instance, China was rated second on the Stanford Human-Centered Artificial Intelligence (HAI) Institute's 2021 Global AI Vibrancy Tool. Stanford University Human-Centered Artificial Intelligence (HAI), "Global AI Vibrancy Tool: Who's Leading the Global AI Race?" accessed October 11, 2023, https://aiindex.stanford.edu/vibrancy/.

[2] Refer to "Scope of Report" for our definition of frontier AI risks.

[3] Kwan Yee Ng et al., "Translation: Artificial Intelligence Law, Model Law v. 1.0 (Expert Suggestion Draft) – Aug. 2023," *DigiChina* (blog), August 23, 2023, https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/.

international cooperation on AI.[4] The Chinese government has also indicated interest in maintaining human control over AI systems and preventing their misuse by extremist groups. However, successful cooperation with China on AI safety hinges on selecting the right international fora for exchanges, as China has expressed a clear preference for holding AI-related discussions under the aegis of the UN.

Technical research in China on AI safety has become more advanced in just the last year. Numerous Chinese labs are conducting research on AI safety, albeit with varying degrees of focus and sophistication. Chinese labs predominantly employ variants of reinforcement learning from human feedback (RLHF) techniques for specification research and have conducted internationally notable research on robustness. Some Chinese researchers have also developed safety evaluations for Chinese Large Language Models (LLMs), although they do not focus on dangerous capabilities. Additionally, several have extensively explored interpretability, particularly for computer vision. While this work diverges in certain aspects from research popular in leading AI labs based in the United States (US) and United Kingdom (UK), the surge in preprint research on AI safety by at least thirteen notable Chinese labs over the past year underscores the escalating interest of Chinese scientists.

Expert discussions around frontier AI risks have become more mainstream in the last year. While some leading Chinese experts expressed worries about risks from advanced AI systems as early as 2016, this was more the exception than the norm. The release of GPT-3 in 2020 spurred more academics to discuss frontier AI risks, but the topic was not yet common enough to merit dedicated discussion in China's top two AI conferences, the World Artificial Intelligence Conference (WAIC) and Beijing Academy of Artificial Intelligence (BAAI) Conference. In 2023, however, frontier AI risks have become a common topic of debate, with multiple Chinese experts signing the Future of Life Institute (FLI) and Center for AI Safety (CAIS) open letters on frontier AI, and the 2023 Zhongguancun (ZGC) Forum and BAAI Conference featuring in-depth discussions on the matter. Several leading experts have also emphasized the Chinese concept "bottom-line

[4] "Foreign Ministry Spokesperson's Remarks on the Global AI Governance Initiative," Ministry of Foreign Affairs (外交部), October 18, 2023, https://www.fmprc.gov.cn/eng/xwfw_665399/s2510_665401/202310/t20231018_11162874.html.

thinking" (底线思维), which bears similarities to the precautionary principle in EU policymaking and offers a unique contribution to explorations on AI risks.

Chinese labs have largely adopted a passive approach to self-governance of frontier AI risks. While numerous labs began releasing ethics principles for AI development in 2018, these were fairly general and did not specifically address the safety of frontier models. More recent action in 2023 by a Chinese AI industry association indicates interest in AI alignment and safety/security issues. Some Chinese labs have publicized safety measures undertaken for their released LLMs, including alignment measures such as RLHF used for models published in 2023. However, the evaluations these labs have publicly stated they conducted primarily focused on truthfulness and toxic content, rather than more dangerous capabilities.

There is a significant lack of data regarding the Chinese public's views of frontier AI. Existing public opinion surveys are outdated, have limited participation, and often lack precise survey questions. However, existing evidence weakly suggests that the Chinese public generally thinks that benefits from AI development outweigh the harms. One survey suggests that the Chinese public and AI scholars do think there are existential risks from Artificial General Intelligence (AGI), but still think AGI should be developed, suggesting that they think the risks are controllable. However, a more comprehensive exploration is essential to understand the Chinese public's views on the significance of frontier AI risks and how to address such risks.

As this is the first report the authors are aware of that seeks to comprehensively map out the AI safety landscape in China, we see it as part of a larger, essential conversation on how China and the rest of the world should act to reduce the increasingly dangerous risks of frontier AI advancement. We hope that this will encourage other institutions to also better our common understanding of AI safety developments in China, which we believe will be beneficial to global security and prosperity.

Table of Contents

Introduction

In April 2023, the 24 highest-ranking officials in China–the Politburo of the Communist Party of China (CPC)—convened a meeting on economic issues. Buried in the meeting readout was a phrase previously alien to public discussions by high officials—Artificial General Intelligence (AGI) (通用人工智能).[5] The full quote read: "[China] must attach importance to the development of artificial general intelligence, create an ecosystem for innovation, and attach importance to the prevention of risks."[6] This announcement revealed that attention to the advantages of but also risks posed by AGI had penetrated the highest levels of government, and also sent a signal to lower-level officials that AGI was an issue worth watching.

Within a month of the Politburo meeting, the Beijing municipal government issued two policies to "promote the innovation and development of artificial general intelligence." The first, "Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing," listed measures to increase the supply of compute and data resources, support the R&D of large models, and encourage the widespread adoption of these models.[7] At the same time, the document reflected the Politburo's concerns about AI risks by calling for third-party safety evaluation benchmarks, model safety/security assessments, and more work on "intent alignment" (人类意图对齐) of large models. The second document, the "Beijing Artificial General Intelligence Industry Innovation Partnership Plan," identified specific industry organizations to help boost the availability of compute and

[5] The Chinese expression 通用人工智能 can be translated as "general purpose artificial intelligence" (general purpose AI) or "artificial general intelligence" (AGI). Historically, the terms "strong AI" (强人工智能) and "general purpose AI/AGI" (通用人工智能) have been used in Chinese policy and general discourse to roughly refer to AI with human-level intelligence and that can perform any task that humans can perform. We have chosen to translate this as AGI given the salience of the term in English discourse, though there is not a universal consensus on how to define AGI in either the English-speaking or Chinese-speaking communities.

[6] Politburo of the Communist Party of China (中共中央政治局), "Analyzing and Researching the Present Economic Situation and Economic Work (分析研究当前经济形势和经济工作)," April 28, 2023, https://www.gov.cn/yaowen/2023-04/28/content_5753652.htm.

[7] General Office of Beijing Municipal People's Government (北京市人民政府办公厅), "Notice by the General Office of the Beijing Municipal People's Government on the Publication of the 'Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing' (北京市人民政府办公厅关于印发《北京市促进通用人工智能创新发展的若干措施》的通知)," May 30, 2023, https://www.beijing.gov.cn/zhengce/gfxwj/202305/t20230530_3116869.html.

training data and to develop algorithmic innovations in order to support the development of the AGI industry.[8]

These developments mirror a growing global awareness of the potential benefits and risks of increasingly generalized, autonomous, and capable AI systems. The United Nations Security Council (UNSC) held its first meeting dedicated to AI in July 2023, in which Secretary-General Guterres stated: “We are derelict in our responsibilities to present and future generations” without action to address risks that generative AI’s creators warn are “potentially catastrophic and existential.”[9] A dialogue in May 2023 between UK Prime Minister Rishi Sunak and CEOs of leading AI labs involved discussion of “existential threats.”[10] Moreover, in September 2023, the European Commission tweeted that: “Mitigating the risk of extinction from AI should be a global priority.”[11] Civil society groups have played a significant part in pushing this awareness, such as through a March open letter by the Future of Life Institute (FLI) calling for a six-month pause on training AI systems more powerful than GPT-4, and a May open letter by the Center for AI Safety (CAIS) on prioritizing the risk of extinction from AI. Both letters were signed by prominent AI scientists and industry leaders around the world, including from China.

What are the prospects for coordination[12] between China and the rest of the world in reducing the risks of frontier AI systems? Notably, the US has put AI, along with other emerging technologies, at the center of what it perceives as a strategic competition with China. Former Secretary of Defense Mark Esper contended in an interview with Semafor

[8] Beijing Economics and Informatization Bureau (北京市经济和信息化局), “Beijing Artificial General Intelligence Industry Innovation Partnership Plan (北京市通用人工智能产业创新伙伴计划),” May 19, 2023, https://www.beijing.gov.cn/zhengce/zhengcefagui/202305/t20230524_3111706.html.

[9] United Nations Security Council, “Security Council Seventy-Eighth Year 9381st Meeting Tuesday, 18 July 2023, 10 a.m. New York, S/PV.9381,” United Nations, July 18, 2023, https://documents-dds-ny.un.org/doc/UNDOC/PRO/N23/210/49/PDF/N2321049.pdf?OpenElement. The meeting transcript can also be accessed by clicking on the hyperlink S/PV.9381 on the Security Council Meetings in 2023 page: https://research.un.org/en/docs/sc/quick/meetings/2023.

[10] Alex Hern and Kiran Stacey, “No 10 Acknowledges ‘Existential’ Risk of AI for First Time,” The Guardian, May 25, 2023, sec. Technology, https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak.

[11] European Commission [@EU_Commission], “Mitigating the Risk of Extinction from AI Should Be a Global Priority. And Europe Should Lead the Way, Building a New Global AI Framework Built on Three Pillars: Guardrails, Governance and Guiding Innovation ↓ Https://T.Co/t7UA9rgN1H,” Tweet, X, September 14, 2023, https://twitter.com/EU_Commission/status/1702295053668946148.

[12] In this report, we define “coordination” and “cooperation” not only as active collaboration but also the passive influence entities can exert on one another, such as leading by example or mutual learning from best practices.

Tech that there is an "existential" race between the US and China on AI.[13] Breathless news headlines fan fears about China overtaking the US in AI development.[14] Such narratives often originate from misunderstandings and ideological oversimplifications, such as the belief that China aims for global AI dominance,[15] or that China flagrantly disregards AI risks.[16] We believe that this framing of AI development as a race to be "won" against geopolitical rivals is excessively simplistic, and there is more room for coordination than many may think.

This report is divided into eight sections that provide a comprehensive overview of the AI safety landscape in China. We contend that China's domestic governance system possesses flexible and potent tools which, although they could be, are not currently primarily oriented towards safety of frontier models. Our section on international AI governance explicates China's key positions on the topic. We detail the current state of technical safety research in China, highlighting longstanding and recent work on issues such as evaluations and value alignment. We guide readers through debates between Chinese experts on risks brought on by AGI, illustrating how a growing number of notable figures from academia and AI labs have begun discussing catastrophic AI risks, making it an increasingly prevalent topic of discourse among Chinese AI scholars. We present initial findings on the safety measures adopted by AI research labs in China and on the Chinese public's views on AI risks.

We hope that this report will serve as a valuable resource for foreign experts and researchers interested in AI safety in China. This report shows that safety of increasingly advanced AI systems is a substantive and growing source of concern among varied and influential actors located in China. While practical cooperation on specific issues can be challenging depending on the context, there are areas where cooperation may be feasible.

[13] Jay Solomon, "Ex-Pentagon Chief: US, China Locked in Existential Struggle for AI Dominance | Semafor," Semafor, July 28, 2023, https://www.semafor.com/article/07/27/2023/us-china-existential-struggle-ai-dominance.

[14] Luiza Ch. Savage and Nancy Scola, "'We Are Being Outspent. We Are Being Outpaced': Is America Ceding the Future of AI to China?," POLITICO, July 18, 2019, https://www.politico.com/story/2019/07/18/global-translations-ai-china-1598442; James Vincent, "China Is about to Overtake America in AI Research," The Verge, March 14, 2019, https://www.theverge.com/2019/3/14/18265230/china-is-about-to-overtake-america-in-ai-research; "China Has Won AI Battle with U.S., Pentagon's Ex-Software Chief Says," Reuters, October 11, 2021, sec. Technology, https://www.reuters.com/technology/united-states-has-lost-ai-battle-china-pentagons-ex-software-chief-says-2021-10-11/.

[15] Seán S. Ó hÉigeartaigh et al., "Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance," Philosophy & Technology 33, no. 4 (December 1, 2020): 571–93, https://doi.org/10.1007/s13347-020-00402-x.

[16] Bill Drexel and Hannah Kelley, "China Is Flirting With AI Catastrophe," Foreign Affairs, May 30, 2023, https://www.foreignaffairs.com/china/china-flirting-ai-catastrophe.

Scope of the Report

In our treatment of AI safety for this report, we focus on the risks from frontier AI outlined by Anderljung et al., namely "The possibility that continued development of increasingly capable foundation models could lead to dangerous capabilities sufficient to pose risks to public safety at even greater severity and scale than is possible with current computational systems."[17] The authors define frontier models as "highly capable foundation models that could exhibit dangerous capabilities." They outline three core problems:

1. **The Unexpected Capabilities Problem.** Dangerous capabilities can arise unpredictably and undetected, both during development and after deployment.
2. **The Deployment Safety Problem.** Preventing deployed AI models from causing harm is a continually evolving challenge.
3. **The Proliferation Problem.** Frontier AI models can proliferate rapidly, making accountability difficult.

These concerns encompass issues like forecasting AI development and safety risks, preventing malicious use of AI systems, and fostering coordination between international actors to resolve the above problems. Consequently, our report encompasses government, expert, and lab statements that touch on risk forecasting, AI misuse, and related topics. Moreover, given that discussion of frontier AI models is often closely intertwined with concepts such as general AI/AGI, Strong AI (强人工智能), and superintelligence (超级人工智能), mentions of these concepts merited inclusion in the paper, where they were more oriented towards safety issues.[18]

The term for AI safety in Chinese (人工智能安全) can mean both AI "safety" and "security," which carry different connotations in English and in academic AI research. This distinction can also be found in Chinese documents. For instance, a report by a government standards organization classifies "security" (外部安全) as "cybersecurity, confidentiality, risk management, physical security, and active defense," whereas "safety" (内部安全) pertains to "control, robustness, reliability, redundancy, and stability."[19] We primarily focus on the

[17] Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety" (arXiv, September 4, 2023), https://doi.org/10.48550/arXiv.2307.03718.

[18] Blaise Agüera Y Arcas and Peter Norvig, "Artificial General Intelligence Is Already Here," *Noema Magazine,* October 10, 2023, https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[19] National AI Standardization Overall Group (国家人工智能标准化总体组) and National Information Technology Standardization Committee AI Subcommittee (全国信标委人工智能分委会), "AI Ethics Governance Standardization Guide, 2023 Version (人工智能伦理治理标准化指南, 2023 版)," March 2023,

"safety" side of the equation and therefore do not analyze issues such as cybersecurity of AI models or vulnerability to theft. Nonetheless, there are times where it is ambiguous whether a Chinese text is referring to AI safety or security, and at times they refer to both. In such uncertain contexts, we translate the term "人工智能安全" as "AI safety/security."

We do discuss AI ethics issues, such as bias, discrimination, and privacy, to some extent in the report, and believe in the importance of addressing such issues in AI systems. However, these topics have received relatively extensive treatment by both international and Chinese academia.[20] Therefore, we discuss AI ethics to the extent that it is important for explaining developments in China that relate to AI safety and where there is overlap between the mechanisms and goals of China's ethics and safety systems. However, we do not more systematically analyze the overall state of AI ethics research in China.

Conflicts of Interest

Concordia AI actively participates in and advises on AI safety within China through various channels, including hosting forums, organizing lectures, and advising on policy.[21] Our ongoing work in this field places us in a unique position to understand and analyze information regarding the state of AI safety in China. However, this engagement also entwines us with the evolving landscape of AI safety in the country, potentially creating a conflict of interest. Our organizational mission and vested interest in advancing AI safety in China might influence our perspective. Nevertheless, we believe our findings are the result of objective analysis, and we disclose this potential conflict to readers for full transparency.

Concordia AI is a social enterprise with a mission to ensure that AI is developed and deployed in a way that is safe and aligned with global interests. In the course of our operations, we have received consulting fees from various companies located in mainland

https://web.archive.org/web/20230531193844/https://www.aipubservice.com/airesource/fs/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BC%A6%E7%90%86%E6%B2%BB%E7%90%86%E6%A0%87%E5%87%86%E5%8C%96%E6%8C%87%E5%8D%97.pdf.

[20] Guangyu Qiao-Franco and Rongsheng Zhu, "China's Artificial Intelligence Ethics: Policy Development in an Emergent Community of Practice," Journal of Contemporary China 0, no. 0 (2022): 1–17, https://doi.org/10.1080/10670564.2022.2153016; Seán S. Ó hÉigeartaigh et al., "Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance," Philosophy & Technology 33, no. 4 (December 1, 2020): 571–93, https://doi.org/10.1007/s13347-020-00402-x; Junhua Zhu, "AI Ethics with Chinese Characteristics? Concerns and Preferred Solutions in Chinese Academia," AI & SOCIETY, October 17, 2022, https://doi.org/10.1007/s00146-022-01578-w.

[21] "Homepage," Concordia AI, accessed October 18, 2023, https://concordia-consulting.com/.

China, Hong Kong, and Singapore, including some discussed in this report. However, no financial engagement with these companies influenced or were related to this report's creation. Concordia AI is an independent institution, not affiliated to or funded by any government or political group.

Domestic AI Governance

Takeaways

- Chinese AI policy has traditionally paid attention to risk prevention from AI, focusing on reliability and controllability.
- 2023 saw a pronounced shift in Chinese policy interest towards AGI, reflected in local policy measures to promote AGI and/or large model development as well as mitigate their risks.
- Existing Chinese regulatory tools—such as the mandatory algorithm registry and safety/security reviews—may serve as groundwork for regulating frontier AI models.
- Beyond regulations, China's voluntary standards, a developing S&T ethics system, third-party certifications, and localized actions could help address frontier AI risks.
- Expert drafts of a forthcoming national AI law already contain provisions that can counteract frontier AI risks, including measures to address high-risk AI scenarios and mandatory reporting of safety incidents.

Introduction

Interest in and focus on AI at the highest levels of the Chinese government began in earnest in 2017. The 2017 Government Work Report[22] was the first work report to reference artificial intelligence, identifying it as a strategic emerging industry that requires greater R&D.[23] Subsequently, the State Council, China's cabinet, published a comprehensive document titled the "New Generation AI Development Plan" (AIDP), which delineated strategic goals for AI development and raised AI development to the level of a national strategy.[24] While the AIDP focuses primarily on AI development, it also mentions the necessity of paying close attention to the risks brought on by AI, increasing ability to foresee and prevent those risks, and ensuring AI safety/security, reliability, and controllability. Thus,

[22] The government work report is annually delivered by the Chinese Premier, officially the head of the Chinese government, during the most important government meeting of the year (the "Two Sessions").

[23] Li Keqiang, "Report on the Work of the Government," The State Council, March 16, 2017, https://english.www.gov.cn/premier/news/2017/03/16/content_281475597911192.htm.

[24] Graham Webster et al., "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)," *DigiChina* (blog), August 1, 2017, https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

early on, the Chinese government was paying attention to risks brought on from AI, with some reference to controllability of AI.

Following the release of the New Generation AI Development Plan, a series of regulations, ethical principles, technical standards, and more were introduced. In June 2019, as part of governance principles published by an expert committee under the Ministry of Science and Technology (MOST), the idea of "agile governance" (敏捷治理) was promulgated, calling for continued research on AI risk and updating of AI governance to promote innovation and account for AI risks.[25] This principle best encapsulates the aspiration of the Chinese government to adapt AI research and regulation in accordance with actual conditions. By 2023, there was a discernible shift in focus toward AGI, likely influenced by the rapid global development of the AI industry. In April 2023, AGI was referenced in a Politburo meeting readout for the first time.[26] The meeting readout underscored the necessity for China to "attach importance to the development of AGI, create an innovative ecology, and emphasize preventing risks."[27] Although this was a short statement nestled within a multi-paragraph readout on promoting the Chinese economy, this first explicit mention of AGI in such a context clearly demonstrates high-level attention to frontier AI development and governance. Policies released by the Beijing municipal government within a month on promoting AGI development further highlight interest in frontier AI.[28]

With AI technology advancing rapidly, China's domestic AI governance system has swiftly adapted over the years. Analyzing this governance entails understanding its multifaceted structure:

[25] Graham Webster and Lorand Laskai, "Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI,'" *DigiChina* (blog), June 17, 2019, https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/.

[26] Politburo of the Communist Party of China (中共中央政治局), "Analyzing and Researching the Present Economic Situation and Economic Work (分析研究当前经济形势和经济工作)," Chinese Government, April 28, 2023, https://www.gov.cn/yaowen/2023-04/28/content_5753652.htm.

[27] "要重视通用人工智能发展，营造创新生态，重视防范风险。"

[28] General Office of Beijing Municipal People's Government (北京市人民政府办公厅), "Notice by the General Office of the Beijing Municipal People's Government on the Publication of the 'Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing' (北京市人民政府办公厅关于印发《北京市促进通用人工智能创新发展的若干措施》的通知)," May 30, 2023, https://www.beijing.gov.cn/zhengce/gfxwj/202305/t20230530_3116869.html; Beijing Economics and Informatization Bureau (北京市经济和信息化局), "Beijing Artificial General Intelligence Industry Innovation Partnership Plan (北京市通用人工智能产业创新伙伴计划)," May 19, 2023, https://www.beijing.gov.cn/zhengce/zhengcefagui/202305/t20230524_3111706.html.

- **Binding Regulations and Laws:** In response to top-level directives, the national administrative state has introduced regulations on specific types of AI: recommender algorithms, deep synthesis, and generative AI. These regulations are led by the Cyberspace Administration of China (CAC). While the regulatory focus has been on content control, social stability, and data security, the new regulatory tools created by these regulations, such as the algorithm registry and safety/security assessments, could be applied to regulating frontier AI risks.
- **Voluntary Standards:** China has begun formulating a voluntary standards system to guide AI safety and security development. Typically, these standards are intended to assist in implementing regulations, like those promulgated by CAC. While existing standards do not directly address frontier AI risks, the most recent white paper by the top national AI standards body hints at concerns like alignment and controllability. There have also been relevant standards on issues such as watermarking and machine learning security.
- **S&T Ethics System:** MOST is in the early stages of constructing a separate S&T ethics system. While the system is *de jure* mandatory, enforcement appears infrequent thus far, and its progress is difficult to verify. It primarily addresses ethical issues, with some references to controllability. While decentralized, it might incorporate safety concerns in the future.
- **Certifications:** Third-party certifications are a form of industry self-governance, where third parties that are sometimes government-affiliated test systems for attributes like safety. No certifications are currently mandated by the government for AI safety, but they could play a future role in ensuring safety of AI systems in China.
- **Local Government Action:** Beyond central level ministries and standards bodies, local governments have initiated policies promoting, and, in some cases, regulating frontier AI applications such as AGI and large models. This suggests some local actors recognize the need for governance in this sphere and provides indications of potential directions for national frontier AI policy.

This section of the report will explore each of the categories listed above, trace China's policy evolution in this arena, and highlight the implications for AI safety. The section will conclude by describing the discussions that will likely occur in the ongoing development of China's national AI Law, and how that will be shaped by the existing governance frameworks.

Binding Regulations and Laws

China's AIDP set forth objectives for China to begin developing local/sectoral regulations by 2020, begin establishing AI laws and regulations by 2025, and construct more comprehensive laws and regulations by 2030.[29] The plan also emphasized the need for a "long-term focus on the impact on social ethics, to ensure that the development of AI falls with [*sic*] the sphere of [being] safe/secure and controllable."[30] These directives set the stage for China to pursue

[29] Graham Webster et al., "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)," *DigiChina* (blog), August 1, 2017, https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

[30] Ibid. In the context of AI, the expression "safe/secure, reliable, and controllable" (安全、可靠、可控) has been used to refer to keeping humans in the loop, ensuring AI systems work as intended, and generally preventing risks from AI. For example, the State Council, in the 2017 New Generation Artificial Intelligence Development Plan wrote: "Uncertainty in the development of artificial intelligence brings new challenges. Artificial intelligence is a disruptive technology with wide impact. It may bring about problems such as changing the employment structure, impacting laws and social ethics, invading personal privacy, and challenging the norms of international relations. It will have far-reaching consequences for government management, economic security, social stability, and even global governance. While vigorously developing artificial intelligence, we must attach great importance to possible security risk challenges, strengthen forward-looking prevention and restraint guidance, minimize risks, and ensure the safe, reliable, and controllable development of artificial intelligence" (人工智能发展的不确定性带来新挑战。人工智能是影响面广的颠覆性技术，可能带来改变就业结构、冲击法律与社会伦理、侵犯个人隐私、挑战国际关系准则等问题，将对政府管理、经济安全和社会稳定乃至全球治理产生深远影响。在大力发展人工智能的同时，必须高度重视可能带来的安全风险挑战，加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展). This quote is about preventing potential risks from AI, safeguarding economic security and social stability, and managing AI's impacts on a range of ethical, social, and governance issues. State Council, "Notice by the State Council on the Publication of the New Generation AI Development Plan (国务院关于印发新一代人工智能发展规划的通知)," July 20, 2017, https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. In a 2018 Politburo study session on AI, President Xi Jinping stated "it is necessary to strengthen the analysis and prevention of potential risks in the development of artificial intelligence, safeguard people's interests and national security, and ensure that artificial intelligence is safe, reliable, and controllable" (要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控). From this statement, the expression "safe, reliable, and controllable" is intimately linked to forecasting and preventing potential risks from AI, and safeguarding public interests and national security. "Xi Jinping: Promote the Healthy Development of China's New Generation AI (习近平：推动我国新一代人工智能健康发展)," Xinhuanet, October 31, 2018, https://www.xinhuanet.com/politics/leaders/2018-10/31/c_1123643321.htm. In the 2023 BRICS Leaders' meeting, President Xi said in his speech: "It is necessary to further expand cooperation in artificial intelligence, strengthen information exchange and technical cooperation, jointly prevent risks, promote the establishment of an international mechanism with universal participation, form a governance framework and standards with broad consensus, and continuously improve the safety, reliability, controllability, and fairness of artificial intelligence technology" (进一步拓展人工智能合作，加强信息交流和技术合作，共同做好风险防范，推动建立普遍参与的国际机制，形成具有广泛共识的治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性、公平性). "Xi Jinping's Speech at the 15th BRICS Leaders Summit - Full Document (习近平在金砖国家领导人第十五次会晤上的讲话 - 全文)," Xinhuanet, August 23, 2023, https://www.news.cn/politics/leaders/2023-08/23/c_1129819257.htm. This statement supports ensuring the safety, reliability, and control over AI, along with AI governance frameworks and standards, by the international community.

regulations on AI. From approximately 2017 to 2020, hard AI regulations were confined to narrow application areas, such as drones and automated driving, seeking to accommodate business in order to foster innovation.[31]

However, the changing relationship between the Chinese government and China's big tech companies, coupled with public concern against the misuse of AI systems, led to a shift toward broader and stricter AI regulations over the course of 2020 and 2021.[32] This shift culminated in three significant Chinese AI regulations: the Administrative Provisions on Algorithm Recommendation for Internet Information Services ("recommendation algorithm regulations"), introduced in draft in 2021 and implemented in 2022; the Provisions on Management of Deep Synthesis in Internet Information Service ("deep synthesis regulations"), drafted in 2022 and implemented in 2023; and the Interim Measures for the Management of Generative Artificial Intelligence Services ("generative AI regulations"), drafted in April 2023 and came into force in August 2023 (refer to Appendix A).[33] These regulations introduced two key regulatory instruments—an algorithm registry and safety/security reviews (some of which are pre-deployment)—which offer a foundation for regulating more advanced, generally capable frontier models.

Therefore, our report takes the expression "safe/secure, reliable, and controllable" to indicate desire to generally prevent risks from AI, ensure that AI systems work as intended, and maintain human control over AI systems.

[31] "Reframing AI Governance: Perspectives from Asia," Digital Futures Lab | Konrad-Adenauer-Stiftung, July 2022, 74–76, https://assets.website-files.com/62c21546bfcfcd456b59ec8a/62df3bbcd1d3f82534a706f1_%E2%80%A2Report_AI_in_Asia.pdf.

[32] "Reframing AI Governance: Perspectives from Asia," Digital Futures Lab | Konrad-Adenauer-Stiftung, July 2022, 78–82, https://assets.website-files.com/62c21546bfcfcd456b59ec8a/62df3bbcd1d3f82534a706f1_%E2%80%A2Report_AI_in_Asia.pdf; "China's Big Tech Crackdown: A Complete Timeline," The China Project, accessed October 11, 2023, https://thechinaproject.com/big-tech-crackdown-timeline/; Rogier Creemers et al., "Is China's Tech 'Crackdown' or 'Rectification' Over?," *DigiChina* (blog), January 25, 2023, https://digichina.stanford.edu/work/is-chinas-tech-crackdown-or-rectification-over/.

[33] Rogier Creemers, Graham Webster, and Helen Toner, "Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022," *DigiChina* (blog), January 10, 2022, https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/; Cyberspace Administration of China (网信办), Ministry of Industry and Information Technology (工信部), and Ministry of Public Security (公安部), "Provisions on Management of Deep Synthesis in Internet Information Service (互联网信息服务深度合成管理规定)," November 25, 2022, https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm; China Law Translate, "Interim Measures for the Management of Generative Artificial Intelligence Services," *China Law Translate* (blog), July 13, 2023, https://www.chinalawtranslate.com/generative-ai-interim/.

The recommendation algorithm regulations represented China's first broad and cross-sectoral (as opposed to industry-specific) regulation impacting AI by regulating algorithms. This regulation created a new tool to do this: an algorithm registry, which Chinese authorities can use to monitor and govern AI algorithms. AI recommendation algorithm providers were now required to start filing their algorithms with the CAC. Furthermore, safety/security assessments were required for algorithms which promote recommendation functions that could "shape public opinion" or "mobilize society."[34] Registrants also had to provide information on datasets used to train the model and upload an algorithm safety/security self-assessment. While many requirements for information disclosure and self-assessment remain missing or ambiguous,[35] given the regulation's emphasis on controlling the effects of recommendation algorithms on social order, the safety/security self-assessment may align with requirements from the 2018 "Regulations for the Security Assessment of Internet Information Services Having Public Opinion Properties or Social Mobilization Capacity" as well as a 2023 standard on machine learning algorithm security.[36] These self-assessment procedures call for, among other things, reporting on the management structures and technical measures to ensure safety/security, as well as basic information on the service provider's services and facilities.

Later regulations on deep synthesis systems ("deepfakes") and generative AI services also require developers and providers to make a filing to the algorithm registry and conduct safety/security self-assessments.[37] While neither the deep synthesis regulations nor

[34] "Translation: New Rules Target Public Opinion and Mobilization Online in China," *DigiChina* (blog), November 21, 2018, https://digichina.stanford.edu/work/new-rules-target-public-opinion-and-mobilization-online-in-china-translation/.

[35] Matt Sheehan and Sharon Du, "What China's Algorithm Registry Reveals about AI Governance," Carnegie Endowment for International Peace, December 9, 2022, https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606.

[36] "Translation: New Rules Target Public Opinion and Mobilization Online in China," *DigiChina* (blog), November 21, 2018, https://digichina.stanford.edu/work/new-rules-target-public-opinion-and-mobilization-online-in-china-translation/; Concordia AI, "AI Safety in China #3," Substack newsletter, *AI Safety in China* (blog), September 20, 2023, https://aisafetychina.substack.com/p/ai-safety-in-china-3.

[37] Note that the deep synthesis regulation was primarily informed by concerns around deepfakes and was finalized preceding the release of GPT-3.5. While its scope effectively covered what many would understand to be "AI generated content," it was not initially written with that conception in mind, and therefore it was followed shortly thereafter with a regulation explicitly on generative AI. China Law Translate, "Provisions on the Administration of Deep Synthesis Internet Information Services." Rogier Creemers and Graham Webster, "Translation: Internet Information Service Deep Synthesis Management Provisions (Draft for Comment) – Jan. 2022," *DigiChina* (blog), February 4, 2022, https://digichina.stanford.edu/work/translation-internet-information-service-deep-synthesis-management-provisi

generative AI regulations shed much further light into the safety/security assessment process, referring instead to their antecedent regulations, they do introduce other safety-relevant requirements. For instance, both regulations mandate the labeling or watermarking of deep synthesis and AI-generated content respectively to address concerns around disinformation.

Overall, these three regulations primarily address a combination of social problems, including content control, privacy, algorithmic discrimination, and anti-competitive practices, and have not yet trained their attention on AI safety-related concerns. However, the existing regulatory capacity and procedural protocols around AI regulation would enable China to seamlessly integrate AI safety concerns into existing or future regulation. For instance, if regulators became concerned about risks of AI self-replication or deception, or AI's ability to generate biological weapons, those concerns could fairly easily be folded into the existing algorithm registry and safety/security assessment system.

Voluntary Standards

China, like the US and European Union (EU), employs standards to guide AI development, and discussions have also occurred around using standards to guide safety of AI systems.[38] In China, standards are often written to help organizations better implement nationwide regulations and are initially introduced as "voluntary." Yet, as regulators frequently cite or recommend them within regulations, these standards can evolve into *de facto* requirements. This places these standards in a gray area between hard law (legally binding regulations) and soft law (such as non-binding guidelines or recommendations). Such an approach exemplifies China's iterative regulatory strategy: testing or encouraging compliance with new standards before potentially codifying them into law. This method allows for flexibility and real-world testing of standards before potential legal codification.

China's AI safety standards system has thus far been primarily focused on cybersecurity and privacy. For example, a broad, authoritative plan on Chinese AI standards, developed by the Standardization Administration of China (SAC), and four other government ministries,

ons-draft-for-comment-jan-2022/. China Law Translate, "Interim Measures for the Management of Generative Artificial Intelligence Services."

[38] "AI Risk Management Framework," NIST, accessed October 11, 2023, https://www.nist.gov/itl/ai-risk-management-framework; Mark McFadden et al., "Harmonising Artificial Intelligence: The Role of Standards in the EU AI Regulation" (Oxford Information Labs, December 2021), https://oxil.uk/publications/2021-12-02-oxford-internet-institute-oxil-harmonising-ai/Harmonising-AI-OXIL.pdf.

merged the section on AI security with privacy protection.[39] Beyond planning documents, China's standardization agencies and research institutes have released numerous white papers analyzing China's AI standards and international trends. These white papers tend to highlight bias, privacy, data security, and cybersecurity. For example, a 2023 white paper from the National Information Security Standardization Technical Committee of China (TC260) enumerates only four AI safety/security standards under development,[40] while white papers by the China Electronic Standardization Institute (CESI) tend to give short shrift to safety/security by merging it in the category with privacy protection.[41]

However, early standards white papers hint at concerns related to frontier AI risks. For instance, CESI's Artificial Intelligence Standardization White Paper released in 2018 states that "AI systems that have a direct impact on the safety of humanity and the safety of life, and may constitute threats to humans" must be regulated and assessed, suggesting a broad threat perception (Section 4.5.7).[42] In addition, a TC260 white paper released in 2019 on AI safety/security worries that "emergence" (涌现性) by AI algorithms can exacerbate the black box effect and "autonomy" can lead to algorithmic "self-improvement" (Section 3.2.1.3).[43]

[39] Standardization Administration of China (国家标准委) et al., "Notice by Five Departments on the Publication of the 'Standardization Construction Guide for National New Generation AI' (五部门关于印发《国家新一代人工智能标准体系建设指南》的通知)," July 27, 2020, https://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm, page 5.

[40] National Information Security Standardization Technical Committee's (TC260) Big Data Security Special Working Group (全国信息安全标准化技术委员会大数据安全标准特别工作组), "AI Safety/Security Standardization White Paper - 2023 Version (人工智能安全标准化白皮书 - 2023 版)," May 2023, https://www.tc260.org.cn/front/postDetail.html?id=20230531105159.

[41] E.g. "Translation: Artificial Intelligence Standardization White Paper (2021 Edition)," Center for Security and Emerging Technology, October 21, 2021, https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper-2021-edition/. CESI is a public institution overseen by the Ministry of Industry and Information Technology (MIIT): "English Introduction (英文介绍)," China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院), accessed October 11, 2023, https://www.cc.cesi.cn/english.aspx.

[42] "Translation: Artificial Intelligence Standardization White Paper," Center for Security and Emerging Technology, May 12, 2020, https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper/.

[43] "Translation: Artificial Intelligence Security Standardization White Paper," Center for Security and Emerging Technology, accessed October 11, 2023, https://cset.georgetown.edu/publication/artificial-intelligence-security-standardization-white-paper-2019-edition/; National Information Security Standardization Technical Committee's (TC260) Big Data Security Special Working Group (全国信息安全标准化技术委员会大数据安全标准特别工作组), "AI Safety/Security Standardization White Paper - 2019 Version (人工智能安全标准化白皮书 - 2019 版)," October 2019, http://www.cesi.cn/images/editor/20191101/20191101115151443.pdf.

Preliminary developments signal that future Chinese AI standards development could address frontier AI risks. The National AI Standardization Overall Group's 2023 guide to ethics governance makes multiple references to human-machine value alignment (人机价值对齐) and calls for pursuing greater standardization involving AI alignment.[44] The guide's view is that alignment involves getting human values into machines through coding, interaction, or pre-learning. It deems alignment crucial for developing certain digital services products and also for achieving strong AI or AGI (Section 5.2.2). While the guide does not list any standards as currently under development regarding alignment, such documents are likely to influence lower-level standards-setting bodies. China also recently began drafting its first standard on large models when the Overall Group announced the creation of a working group on large models at the WAIC in Shanghai in July 2023, jointly led by Shanghai AI Lab's **QIAO Yu (乔宇)** and representatives from six other Chinese tech giants.[45] The creation of this group indicates concern about and shows responsiveness to the explosion of generative AI products over the past year in China. However, the standard's drafting remains in an early stage, and the specific risks it will address remain uncertain.

S&T Ethics System

China has set up an S&T ethics system under MOST that includes AI and requires researchers to undergo some level of review and scrutiny in the development process. However, the system's decentralized nature, coupled with ambiguous enforcement and implementation levels, poses challenges. Nevertheless, it attempts to incentivize AI developers to align their systems with ethical principles, primarily focusing on bias and privacy, with some attention to human control.

[44] Following the New Generation AI Development Plan's calls for China to establish an AI technology standards and IP system, in January 2018, the "National AI Standardization Overall Group" and "Expert Consultation Group" (国家人工智能标准化总体组和专家咨询组) were created to lead and coordinate development of AI standards. "National AI Standardization Overall Group and Expert Consultation Group Created (国家人工智能标准化总体组和专家咨询组成立)," China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院), January 19, 2018, http://www.cesi.cn/201801/3539.html; National AI Standardization Overall Group (国家人工智能标准化总体组) and National Information Technology Standardization Committee AI Subcommittee (全国信标委人工智能分委会), "AI Ethics Governance Standardization Guide, 2023 Version (人工智能伦理治理标准化指南, 2023 版)," March 2023, https://web.archive.org/web/20230531193844/https://www.aipubservice.com/airesource/fs/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BC%A6%E7%90%86%E6%B2%BB%E7%90%86%E6%A0%87%E5%87%86%E5%8C%96%E6%8C%87%E5%8D%97.pdf.

[45] "Shanghai AI Lab Selected as the Leader of the National AI Standardization Overall Group's Large Model Special Group 上海人工智能实验室当选国家人工智能标准化总体组大模型专题组组长," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), accessed October 11, 2023, https://www.shlab.org.cn/news/5443434.

China's work on AI ethics began in earnest under MOST's National New Generation AI Governance Expert Committee (国家新一代人工智能治理专业委员会), which in September 2021 released "Ethical Norms for New Generation AI."[46] The few references to AI safety/security in the document occurred primarily in the context of privacy and data security. However, it did provide some focus on ensuring that "AI is always under human control" under the principle of "controllability," calling for AI to "abide by shared human values" (Article 3). The document also called for enhancing security and transparency, emphasizing explainability and controllability, increasing AI's resistance to interference, and improving auditability (Article 12).

In March 2022, the CPC Central Committee General Office and State Council General Office published "Opinions on Strengthening Science and Technology Ethics Governance."[47] This document puts greater government heft behind China's S&T ethics apparatus, noting that AI is also a "key area" for ethics (Section 4.1, 4.3). It codifies the leading role of the National Committee for S&T Ethics (国家科技伦理委员会) under MOST (Section 3.1) and establishment of S&T ethics review committees by universities, research institutes, and companies (Section 3.2). Initial governance of S&T ethics will focus on a "high-risk" activities list drawn up by the National Committee for S&T Ethics (Section 5.2).

In October 2023, MOST and nine other government ministries released the "Science and Technology Ethics Review Plan (Trial)," which will be effective starting December 1, 2023.[48] While the document would not explicitly require S&T ethics reviews for most AI-related

[46] "Translation: Ethical Norms for New Generation Artificial Intelligence Released," Center for Security and Emerging Technology, October 21, 2021, https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/; "'Ethical Norms for New Generation AI' Published (《新一代人工智能伦理规范》发布)," Ministry of Science and Technology (科技部), September 26, 2021, https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

[47] "General Office of the Communist Party of China and General Office of the State Council on Publishing 'Opinions On Strengthening Science and Technology Ethics Governance' (中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》)," March 20, 2022, https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm; Yi Zeng (曾毅), "Opinion on Strengthening the Ethics and Governance in Science and Technology," International Research Center for AI Ethics and Governance, March 22, 2022, https://ai-ethics-and-governance.institute/2022/03/22/china-released-opinion-on-strengthening-the-ethics-and-governance-in-science-and-technology/.

[48] Ministry of Science and Technology (科技部) et al., "Notice on the Publishing of the 'Science and Technology Ethics Review Plan (Trial)' (关于印发《科技伦理审查办法（试行）》的通知)," October 8, 2023, https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html.

research, it does require institutions conducting AI research that involves "sensitive fields of ethics" to create an S&T ethics review committee. Most ethics reviews would be conducted by the review committee, but for applications on a special "expert review" list published by MOST, the local government or ministerial authority will set up a separate review by unaffiliated, outside experts before developing the technology. AI applications on the "expert review" list include, at present: human-machine integrations that have a strong effect on human emotions and health, algorithms that have social or public mobilization characteristics, and autonomous decision-making systems that are highly autonomous and pose risks to safety or personal health. This encompasses a significant and relatively broad category of AI systems, and requiring reviews during the R&D phase is unusual among other Chinese regulations.

Overall, the S&T ethics system is still in the early phase of construction, and implementation is decentralized, with separate ethics review boards in each relevant university or research institution. While institutions like Tsinghua University and Alibaba have begun implementing AI ethics reviews, it is difficult to track the actual pace of implementation, and hard to tell how stringently the review committees are conducting evaluations.[49] There is little evidence either way on whether AI ethics review processes have meaningfully changed developer behavior. This system has yet to play a substantial role in addressing frontier risks, given the limited mentions of risks related to human control among broader issues like privacy, discrimination, and bias. As this system evolves, it might offer a platform for evaluating frontier risks during development. However, it lacks the policy force of the algorithm registry and safety/security assessment system.

Certifications

Third-party certifications could play a role in ensuring safety of AI systems in China. This area remains under-researched, as few scholars have studied China's certifications on cybersecurity, data security, and related cyber issues. Certifications constitute a form of industry self-governance, in which third parties—sometimes government-affiliated, sometimes independent—test an organization's capabilities and provide them with some

[49] "Tsinghua University Establishes a Science and Technology Ethics Committee (清华大学科技伦理委员会成立)," Tsinghua University (清华大学), December 31, 2022, https://www.tsinghua.edu.cn/info/1177/100966.htm; "Alibaba Group CTO Cheng Li: the science and technology ethics governance committee should be the 'gatekeeper' of technological innovation (阿里集团CTO程立：科技伦理治理委员会要做技术创新的'守门人')," China News (中新网), September 2, 2022, https://www.sh.chinanews.com.cn/kjjy/2022-09-02/102917.shtml.

sort of certificate if they pass. Thus far, certifications are not mandated by the government in the field of AI safety, though they could eventually become a more official regulatory tool, as it has become for personal information protection.[50] MOST's S&T ethics review plan, published in October 2023, previews that it plans to promote development of a certifications system for S&T ethics reviews.[51]

One existing example is a large model testing system that is being developed by the China Academy of Information and Communications Technology (CAICT). This system aims to assess safety and trustworthiness, though the safety component remains unspecified.[52] CAICT also developed a test for "trustworthy AI" that it was implementing by at least 2021.[53] Additionally, many of CAICT's certifications are developed in accordance with the Artificial Intelligence Industry Alliance (AIIA).[54]

At present, the specifics of these certification processes remain unclear, as does their role in AI safety governance. Some details on CAICT's tests are public, such as the involvement of a lab under the Ministry of Industry and Information Technology (MIIT), and that at least 90 companies have been tested.[55] However, most details are difficult to access. Third-party certifications might emerge as a viable path for developing independent evaluations of AI

[50] "Personal Information Protection Certification Implementation Rules (个人信息保护认证实施规则)," Cyberspace Administration of China (网信办), November 18, 2022, http://www.cac.gov.cn/2022-11/18/c_1670399936983876.htm.

[51] Ministry of Science and Technology (科技部) et al., "Notice on the Publishing of the 'Science and Technology Ethics Review Plan (Trial)' (关于印发《科技伦理审查办法（试行）》的通知)," October 8, 2023, https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html, Article 41.

[52] Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Trustworthy AI Technology Hotspot | Large Models Continually Release Technological Dividends, and an Evaluation System for the Extremely Large Model Industry Is Formally Released (可信AI技术热点｜大模型持续释放技术红利，产业级大模型评估体系正式发布)," Weixin Official Accounts Platform, June 27, 2022, http://mp.weixin.qq.com/s?__biz=MzU0MTEwNjg1OA==&mid=2247499125&idx=2&sn=fc677dcdd56cc78b59563798bfedc2c7&chksm=fb2c4ab0cc5bc3a687deebc43d07a53b3e0ac79829fc77a4b8cbd4630824c622c2191005dbf2#rd.

[53] "CAICT Formally Begins Second Batch of 'Trustworthy AI' Evaluations for 2021 (中国信通院2021年第二批'可信AI'评测正式启动)," China Academy of Information and Communications Technology (CAICT) (中国信通院), September 23, 2021, http://www.caict.ac.cn/xwdt/ynxw/202109/t20210923_390249.htm.

[54] "Services (联盟服务)," Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), March 29, 2019, http://aiiaorg.cn/index.php?m=content&c=index&a=show&catid=34&id=58.

[55] Trustworthy AI Evaluations (可信AI评测), "CAICT's 8th Batch of 'Trustworthy AI' Evaluations in 2023 Formally Begins (中国信通院2023年'可信AI'（第八批）评测正式启动)," Weixin Official Accounts Platform, February 17, 2023, http://mp.weixin.qq.com/s?__biz=Mzg3ODU5NDI0MQ==&mid=2247487529&idx=2&sn=eeef8e2f145ccb8ee8e248bec93725b8&chksm=cf100187f86788911a4c0d1264349b0e431d5edb1ef17b56bbbeb4a96935dcae0d1318ca9c8b#rd.

model safety in the future, but they could also merely serve as a way for companies to superficially demonstrate commitment to AI safety without being meaningfully evaluated.

Local Government Action

At the local government level, there have also been several important developments that more directly involve AI safety than most central government efforts thus far. Given China's practice of piloting policies at the local level before rolling them out nationwide, these policies provide early indicators of potential future directions of central government policy. However, there is no guarantee that these specific policies will be mirrored at the central level.

The Beijing municipal government released measures on promoting the development of AGI in May 2023.[56] To our knowledge, this is the first policy on AGI at the provincial level.[57] The overall focus of the measure is to improve Beijing's AGI development capabilities, such as expanding computing power supply, improving training data, and pursuing multiple paths to AGI. At the same time, the measures exhibit multiple positive signals of concern around AGI risks. For instance, the measures call for greater research around human intent alignment, though curiously refers to this as a fine-tuning method, which might suggest underestimation of the difficulty of aligning more generally capable AI systems (Article 7). The document also calls for nonprofits to build model evaluation systems and benchmarks for foundation models, including testing for, primarily, capability factors, but also robustness (Article 9). The measures additionally note support for S&T ethics and security norms, albeit without substantial elaboration (Article 21). Another Beijing government document on AGI industry innovation partnerships in May did not mention AI safety, focusing more on promotion mechanisms.[58] Given that Beijing is arguably the leading locality for LLM development in

[56] General Office of Beijing Municipal People's Government (北京市人民政府办公厅), "Notice by the General Office of the Beijing Municipal People's Government on the Publication of the 'Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing' (北京市人民政府办公厅关于印发《北京市促进通用人工智能创新发展的若干措施》的通知)," May 30, 2023, https://www.beijing.gov.cn/zhengce/gfxwj/202305/t20230530_3116869.html.

[57] Note: Beijing is one of four "directly-administered municipalities" (直辖市) which means that it is treated as a province. "Directly-Administered Municipalities (直辖市)," Baidu Baike (百度百科), accessed October 17, 2023, https://baike.baidu.com/item/%E7%9B%B4%E8%BE%96%E5%B8%82/725471.

[58] Beijing Economics and Informatization Bureau (北京市经济和信息化局), "Beijing Artificial General Intelligence Industry Innovation Partnership Plan (北京市通用人工智能产业创新伙伴计划)," May 19, 2023, https://www.beijing.gov.cn/zhengce/zhengcefagui/202305/t20230524_3111706.html.

China, home to 38 of 79 LLMs documented by a Chinese government think tank in May 2023, this development is especially important.[59]

The Chengdu Municipal Bureau of Economic and Information Technology also released measures on accelerating large model innovation in August 2023.[60] However, these measures were entirely focused on development with no mention of safety, alignment, or risks.

Both Shenzhen and Shanghai municipal governments released regulations more broadly on AI industry development, in August and October 2022 respectively.[61] These regulations, bearing notable similarities, primarily emphasize development, promotion, and application. However, both regulations also call for the creation of municipal AI ethics committee to publish guidelines and white papers on AI ethics and safety and oversee implementation of norms in AI enterprises. Additionally, both regulations highlight increased requirements for high-risk AI products through a pre-deployment testing. Specifically, the Shanghai regulation mandates a review of product controllability and stipulates that biometric technology providers should provide "safe and controllable technical guarantee measures" (Article 68).

In summary, the Beijing and Chengdu measures show how local governments are responding to rapid AI development by focusing on AGI and large models, respectively. The Beijing measures even raise the topic of alignment, albeit with a focus on intent rather than value alignment. The Shenzhen and Shanghai regulations, while broader, single out more intensive

[59] Jia Liu (刘佳), "China Already Has 79 Large Models with 1 Billion Parameters, and Industry Is Calling for Establishing a 'Moat' for Independent Innovation as Quickly as Possible (中国已有79个10亿参数大模型，业界呼吁尽快建立自主创新'护城河')," Yicai (第一财经), May 29, 2023, https://m.yicai.com/news/101769137.html.

[60] Chengdu Municipal Bureau of Economic and Information Technology (成都市经济和信息化局) and Chengdu New Economic Development Commission (成都市新经济发展委员会), "Notice on the Publication of 'Several Measures for Accelerating Large Model Innovation and New Applications to Promote the High Quality Development of the AI Industry in Chengdu Municipality' (关于印发《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》的通知)," Chengdu Municipal Bureau of Economic and Information Technology (成都市经济和信息化局), August 4, 2023, https://cdjx.chengdu.gov.cn/cdsjxw/c160798/2023-08/04/content_ba168b8475ad4419ac6c7393012e2f5c.shtml; "20 Measures! Chengdu Accelerates the Innovation and Application of Large Models to Promote the High Quality Development of the AI Industry (20条措施！成都加快大模型创新应用推进人工智能产业高质量发展)," The Paper (澎湃), August 7, 2023, https://m.thepaper.cn/newsDetail_forward_24145627.

[61] Like Beijing, Shanghai is at the same level as provinces. However, Shenzhen is not. "Shenzhen Special Economic Zone AI Industry Promotion Measures (深圳经济特区人工智能产业促进条例)," Shenzhen Municipal People's Congress (深圳市人大常委会), September 9, 2022, http://www.szrd.gov.cn/szrd_zlda/szrd_zlda_flfg/flfg_szfg/content/post_834707.html; "Regulations for Promoting the Development of the Artificial Intelligence Industry in Shanghai (上海市促进人工智能产业发展条例)," Shanghai Municipal People's Government (上海市人民政府), October 1, 2022, https://www.shanghai.gov.cn/hqcyfz2/20230627/3a1fcfeff9234e8e9e6623eb12b49522.html.

requirements for higher risk AI applications, though the basis for this risk calculation is not made clear. Notably, the Shanghai municipal government's measures emphasize evaluating models based on controllability, signaling regulatory attention to this crucial aspect.

Implications and China's Forthcoming AI Law

In June 2023, China's State Council unveiled the intention to draft a national AI law in its annual legislative plan, marking the beginning of an expected flurry of activity over the next few years to frame, draft, and review China's AI Law proposals.[62] While this law will likely present a broad AI governance strategy rather than delve into specifics, it will mark a pivotal juncture in China's domestic AI governance, shaping future regulations and standards.

A draft model AI Law, written by experts convened by the Chinese Academy of Social Sciences Legal Research Institute's Cyber and Information Law Research Office, offers an initial insight into the direction of China's AI Law.[63] Drafted by non-governmental experts to fuel national AI Law discussions, this model law, though not the final version, reflects Chinese expert viewpoints on AI governance.

The draft law aims to regulate certain AI scenarios through a negative list, necessitating government-approved permits for certain AI research, development, and services. It delineates the roles and responsibilities of AI researchers, developers, and providers, proposing a National AI Office for AI regulation and oversight. Treatment of risks from frontier models falls into several categories.

- **Foundation Models:** The draft stipulates extra compliance requirements for foundation models due to their key role in the industry chain.[64] In particular,

[62] "Announcement by the State Council General Office on Publishing the 2023 State Council Yearly Legislative Plan (国务院办公厅关于印发国务院2023年度立法工作计划的通知)," State council, June 6, 2023, https://www.gov.cn/zhengce/content/202306/content_6884925.htm.

[63] Kwan Yee Ng et al., "Translation: Artificial Intelligence Law, Model Law v. 1.0 (Expert Suggestion Draft) – Aug. 2023," *DigiChina* (blog), August 23, 2023, https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/.

[64] New Governance (新治理), "'AI Law (Model Law) 1.0' (Expert Suggestion Draft) Drafting Statement and Full Document (《人工智能法(示范法)1.0》(专家建议稿)起草说明和全文)," Weixin Official Accounts Platform, August 15, 2023, http://mp.weixin.qq.com/s?__biz=Mzg2NDYzMzMyMA==&mid=2247485974&idx=1&sn=f543b06dc59b3e81dfb0b45ad744b6d1&chksm=ce672291f910ab87083938537f71a4d7b4e313d00365d8a8a69c493b11233388ea23d6c1cb7b#rd.

organizations that pursue foundation model R&D would need to accept yearly "social responsibility reports" by predominantly independent institutions (Article 43).

- **Worries about loss of control:** Preventing loss of control is one criterion for securing a negative list permit. This necessitates staff with strong knowledge of human supervision and technical assurance measures to ensure AI remains safe and controllable (Article 25 Clauses 3 and 5). Providers on the negative list must also guarantee humans can intervene or take over at any point during the automated operation of AI services (Article 50 Clause 4).
- **Government reporting:** Safety incidents will need to be reported to the government (Article 34). Government regulators can also summon companies to meet about safety risks or incidents and can require companies to undergo compliance audits by a professional institution (Article 54).
- **Safety/security and risk assessments:** Various provisions in the draft model law refer to risk management assessments by AI developers or the government, safety/security assessments, and ethics reviews (Article 38, 39, 40, 42, 51). The exact risks and security concerns are not specified, but likely focus on cybersecurity, data security, privacy, and discrimination related issues. However, the drafters appear somewhat concerned about more dangerous capabilities too, given a reference to "emergent" properties of models in the statement accompanying their draft.[65]
- **Technical safety research:** The draft calls for government support to organizations researching and developing technology related to AI monitoring and warning, safety/security assessments, and other regulatory or compliance-related technologies.

The central government of China has primarily focused its efforts to mitigate AI risks on managing public opinion, addressing unfair commercial practices, and preventing data breaches, as can be seen in the S&T ethics system, existing standards, and binding regulations. However, recent months have seen central-level expressions of concern about AGI and frontier AI risks, notably in the April 2023 Politburo meeting and in a guide to ethics governance standards. Meanwhile, the Beijing municipal government has also become publicly interested in AGI development and safety.

[65] New Governance (新治理), "'AI Law (Model Law) 1.0' (Expert Suggestion Draft) Drafting Statement and Full Document (《人工智能法（示范法）1.0》（专家建议稿）起草说明和全文)."

The drafting process of China’s AI Law will undoubtedly see many changes. The direction over the past several months points to increasing government interest in regulating and preventing frontier AI risks, but the final codified law remains uncertain. At a minimum, China’s existing system of strong regulation of AI algorithms and content generation, a growing voluntary standards system, and limited but expanding AI ethics reviews offers tools that could be used in the future to increase safe and secure AI development, even though the focus of these tools is not currently on frontier capabilities and risks.

International AI Governance

Takeaways

- China has become more proactive on the issue of international AI governance in 2023, most notably announcing a Global AI Governance Initiative in October 2023.
- China advocates for inclusive international AI governance, emphasizing UN-led governance and sharing AI benefits with developing nations.
- SAC and CESI have represented China in the ISO AI subcommittee, ISO/IEC JTC 1/SC 42, indicating active engagement in international AI standards.
- China has contributed to AI ethics discussions at the G20, UNESCO, and ITU.
- The Chinese stance underscores the importance of human control over AI.
- China also supports international coordination against AI misuse by terrorists.

Introduction

Since 2016, China has actively participated in international AI governance, especially with regard to discussions surrounding lethal autonomous weapons, ethical principles, and standards-setting. China's initial involvement in international AI governance centered on AI ethics discussions within UN bodies as well as the regulation of lethal autonomous weapons. In December 2016, China submitted its first paper on lethal autonomous weapons (LAWs) to the UN Convention on Certain Conventional Weapons (CCW) and became the first Permanent Five member to advocate for international legislation on LAWs.[66] In ensuing years, it also supported the formulation of the Group of Twenty's (G20) AI Principles, announced in 2019, and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence, launched in 2021. Moreover, China has consistently engaged with the negotiation and formulation of international AI standards through the International Organization for Standardization (ISO) and International Telecommunication Union (ITU). See Appendix B for a list of the international AI governance documents referenced in this section.

However, 2023 appears to have marked a phase shift in China's efforts to participate in international AI governance, underscored by the release of a Global AI Governance Initiative in October 2023. In April 2023, China elevated international cooperation on AI governance

[66] Debates over the Chinese government's definition of LAWs is outside the scope of this report.

within one of the country's flagship foreign policy programs, the Global Security Initiative, which listed the international governance of AI and other emerging technologies as one of twenty "cooperation priorities."[67] China's Ministry of Foreign Affairs' (MOFA) "Proposal of the People's Republic of China on the Reform and Development of Global Governance," published in September 2023, similarly identified AI as a "frontier" for global governance.[68] Then, President Xi Jinping announced the new Global AI Governance Initiative (全球人工智能治理倡议) at the opening ceremony of the Third Belt and Road Forum for International Cooperation in October.[69] The full document sets out China's core positions on values for AI development and areas for international cooperation on AI.[70] These developments were accompanied by an uptick of interest in AI governance from other countries such as the US and UK.[71] Regardless, much of global AI governance remains fractured across geopolitical fault lines, with China having been excluded by default from most of the international AI governance forums such as the Organisation for Economic Co-operation and Development (OECD), Group of Seven (G7), and the Global Partnership on AI (GPAI). As global debates on the governance of AI, and especially of frontier AI risks, have proliferated over the past year, the question of how China is to be involved in international AI governance has become increasingly crucial.

To shed light on this question, this section will first outline three features of China's approach to AI governance over the past six years. Then, it will elaborate on China's views

[67] "The Global Security Initiative Concept Paper," Ministry of Foreign Affairs (外交部), February 21, 2023, https://www.fmprc.gov.cn/mfa_eng/wjbxw/202302/t20230221_11028348.html. The Global Security Initiative was first announced by President Xi in April 2022 at the Boao Forum and is one of three such global initiatives, alongside the Global Development Initiative and the Global Civilization Initiative. The Global Security Initiative reiterates a number of longstanding Chinese foreign policy positions, such as support for territorial sovereignty and the central role of the UN Charter. It also elevates the concept of "indivisible security (不可分割的安全)," which criticizes North Atlantic Treaty Organization (NATO) expansion and US military alliances in East Asia.

[68] "Full Text: Proposal of the People's Republic of China on the Reform and Development of Global Governance," September 13, 2023, https://english.news.cn/20230913/edf2514b79a34bf6812a1c372dcdfc1b/c.html?mc_cid=8031f71d00&mc_eid=ccbfb1d564.

[69] "Foreign Ministry Spokesperson's Remarks on the Global AI Governance Initiative," Ministry of Foreign Affairs (外交部), October 18, 2023, https://www.fmprc.gov.cn/eng/xwfw_665399/s2510_665401/202310/t20231018_11162874.html.

[70] "Global AI Governance Initiative (全球人工智能治理倡议)," Cyberspace Administration of China (网信办), October 18, 2023, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[71] Center for AI Safety, "The Landscape of US AI Legislation," AI Safety Newsletter, September 19, 2023, https://newsletter.safe.ai/p/the-landscape-of-us-ai-legislation; Department for Science, Innovation and Technology, "UK Government Sets out AI Safety Summit Ambitions," GOV.UK, September 4, 2023, https://www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions; Astha Rajvanshi, "Rishi Sunak Wants the U.K. to Be a Key Player in Global AI Regulation," Time, June 14, 2023, https://time.com/6287253/uk-rishi-sunak-ai-regulation/.

relating to maintaining human control over AI systems and preventing misuse of AI by extremist groups, areas that might offer opportunities for broader cooperation. Finally, the section will combine observations of China's approach to international AI governance and Chinese views on AI risks to suggest how, through measures such as agreeing on technical standards and monitoring mechanisms, China and other countries may strengthen their cooperation on common risks from AI.

Features of China's Approach

Calling for Greater Inclusivity in International AI Governance

In recent years, China's leadership has proposed a notion of "true multilateralism," contrasting it with "fake" or "pseudo-multilateralism." This notion primarily emphasizes the central role of the UN, countering the unilateralism commonly practiced by powerful international entities. The second characteristic of "true multilateralism" is its inclusivity, as opposed to "pseudo-multilateralism," which, in Beijing's view, serves as a pretext to create exclusive blocs or divide the world ideologically.[72]

This tenet of recent Chinese foreign policy has manifested in the country's approach to international AI governance. China advocates for a significant UN role in any AI governance efforts. This is stated explicitly in the final paragraph of the Global AI Governance Initiative: "We support discussions within the United Nations framework to establish an international institution to govern AI, and to coordinate efforts to address major issues concerning international AI development, security, and governance."[73] Moreover, MOFA's proposal on global governance calls for "uphold[ing] the international system with the United Nations at its core, [and] support[ing] the UN in playing a central role in international affairs."[74] Furthermore, China's key documents and speeches on international AI governance have

[72] Anna Caffarena, "Why China's Understanding of Multilateralism Matters for Europe," April 2022, http://www.eurics.eu/upload/document/20220427040433_why-chinas-understanding-multilateralism-matters-for-europe-eurics-2022.pdf; Yi Wang (王毅), "Staying Open and Inclusive and Upholding Multilateralism: Toward a Community with a Shared Future for Mankind," Ministry of Foreign Affairs (外交部), May 26, 2021, https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/zyjh_665391/202105/t20210526_9170548.html.

[73] "Global AI Governance Initiative (全球人工智能治理倡议)," Cyberspace Administration of China (网信办), October 18, 2023, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[74] "Full Text: Proposal of the People's Republic of China on the Reform and Development of Global Governance," September 13, 2023, https://english.news.cn/20230913/edf2514b79a34bf6812a1c372dcdfc1b/c.html?mc_cid=8031f71d00&mc_eid=ccbfb1d564.

mostly been expressed through the UN. China's position papers on LAWs were presented to CCW, and Ambassador **ZHANG Jun's (张军)** speech at the UN Security Council was another inflection point in demonstrating a level of Chinese government concern over frontier AI risks. China has also engaged in AI norms and standards-setting processes under UN-subordinated bodies, such as UNESCO and the ITU.

In addition to endorsing the UN's role, China attempts to support the interests of developing countries through advocating for wide access to AI technology and its benefits. This is apparent in the Global AI Governance Initiative's call for all countries to "have equal rights to develop and use AI."[75] This would entail global collaboration to share AI knowledge and make AI technologies available to the public under open-source terms.[76] In addition, the 2022 MOFA position paper on ethical governance of AI, which calls on countries to "ensure that the benefits of AI technologies are shared by all countries."[77] Similarly, the Chinese ambassador to the UN's July 2023 speech suggested that international actors "ensure that developing countries equally enjoy the development dividends brought by AI technology and continuously enhance their representation, voice, and rights of decision-making in this field."[78]

Continued exclusion of China in major international AI governance fora may encourage parallel AI governance efforts. Over the past two years, BRICS has made multiple announcements relating to AI, calling for members to share best practices on ethical and responsible AI in 2022 and creating an AI Study Group in 2023.[79] Notably, the Global AI Governance Initiative was unveiled at the Belt and Road Forum, reflecting China's proactive approach to shape AI governance through its own initiative. While separate international AI governance efforts need not be adversarial, a fragmented landscape among major AI powers

[75] "Global AI Governance Initiative (全球人工智能治理倡议)," Cyberspace Administration of China (网信办), October 18, 2023, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[76] Note that the reference to open-source does not necessarily mean that the Chinese government supports full open-sourcing of all AI research. References in the document to "tiered" management of AI based on "risk levels" suggest otherwise.

[77] "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)," Ministry of Foreign Affairs (外交部), November 17, 2022, https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html.

[78] Jun Zhang (张军), "Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security," Permanent Mission of the People's Republic of China to the UN (中华人民共和国常驻联合国代表团), July 18, 2023, http://un.china-mission.gov.cn/eng/hyyfy/202307/t20230719_11114947.htm.

[79] Concordia AI, "AI Safety in China #2," Substack newsletter, *AI Safety in China* (blog), September 7, 2023, https://aisafetychina.substack.com/p/ai-safety-in-china-2.

is likely to hinder greater alignment of standards and actions crucial to addressing frontier AI risks.

Engagement on International AI Standards

China played a pivotal role in the creation of one of the key international standards bodies for AI. SAC and CESI hosted the first plenary meeting of the ISO AI subcommittee, ISO/IEC JTC 1/SC 42.[80] A Chinese expert was also the first leader of one of the four working or study groups in ISO/IEC JTC 1/SC 42.[81] China has since regularly attended meetings of ISO/IEC JTC 1/SC 42 and is actively engaged in standards development processes at ISO.[82] While it is difficult to quantify the size of China's contribution to international AI standards from publicly available information, in their white papers, Chinese domestic standards institutions regularly reference and summarize ISO standards and national standards by other countries.

Participation in AI Ethics Principles

China has also actively taken part in non-binding efforts to develop AI ethics principles in various international fora. In June 2019, the G20, of which China is a member, adopted the G20 AI Principles.[83] This concise, high-level document advocates for AI to be human-centered, transparent and explainable, robust and secure, and accountable. It also offers policy recommendations for governments to invest in AI R&D, create an agile policy environment, and pursue international cooperation. China also endorsed a UNESCO recommendation on Ethics of AI in November 2021, with one non-governmental Chinese

[80] "ISO/IEC JTC 1/SC 42 - Artificial Intelligence," ISO, accessed October 11, 2023, https://www.iso.org/committee/6794475.html; "Our Institute Undertook the First Plenary Meeting of the ISO/IEC JTC 1/SC 42 AI Technical Sub-Committee (我院承办ISO/IEC JTC 1/SC 42人工智能分技术委员会第一次全会)," China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院), April 23, 2018, http://www.cesi.cn/201804/3821.html.

[81] "Our Institute Undertook the First Plenary Meeting of the ISO/IEC JTC 1/SC 42 AI Technical Sub-Committee (我院承办ISO/IEC JTC 1/SC 42人工智能分技术委员会第一次全会)."

[82] "Our Institute Attended the Third Plenary Meeting of ISO/IEC JTC 1/SC 42 (我院参加ISO/IEC JTC 1/SC 42第三次全体会议)," China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院), May 6, 2019, http://www.cesi.cn/201905/5057.html; Matt Sheehan and Jacob Feldgoise, "What Washington Gets Wrong About China and Technical Standards," Carnegie Endowment for International Peace, February 27, 2023, https://carnegieendowment.org/2023/02/27/what-washington-gets-wrong-about-china-and-technical-standards-pub-89110.

[83] "G20 AI Principles," OECD.ai Policy Observatory, June 9, 2019, https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf.

expert participating in that document's drafting.[84] The UNESCO document, more lengthy and detailed, outlines AI values and principles and offers policy recommendations, including ethical impact assessments, promoting data security, and fostering development. Various Chinese entities have also participated in efforts at the ITU to pursue principles around "AI for Good."[85]

Views Relating to the International Governance of Frontier AI risks

Maintaining Human Control Over AI Systems

The potential for humanity to lose control over increasingly advanced and generally capable AI systems is a key frontier AI concern. Ensuring human control over AI has been a consistent element in China's domestic AI policy documents and also in its expressions on international AI governance. Evidence for the former is outlined in the Domestic AI Governance section of this report; this section will elaborate on the latter.

While the principles of safety, reliability, and controllability of AI systems are frequently reiterated in domestic governance, China's international governance documents initially focused on ensuring human control of weapons systems, specifically. This focus is reflected in MOFA's position papers submitted to the UN CCW. In 2016, amid emerging debates on ensuring human control over LAWs, the Chinese position paper requested greater clarity on the role of humans in LAWs. In a subsequent position 2021 position paper, MOFA provided a concretized stance on the role of human control, calling for "relevant weapon systems [to remain] under human control and efforts must be made to ensure human suspension [of relevant weapons systems] at any time." However, MOFA's 2022 position paper submitted to CCW expanded calls for human control beyond military AI systems to AI systems more broadly. The paper stated that AI should be "safe, reliable, [and] controllable," and called for

[84] UNESCO, "Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library," UNESDOC Digital Library, November 23, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000381137.
[85] "ZTE Showcases Data-Driven Intelligence at ITU AI for Good 2023 Summit," ZTE, July 19, 2023, https://www.zte.com.cn/content/zte-site/www-zte-com-cn/global/about/news/zte-showcases-data-driven-intelligence-at-itu-ai-for-good-2023-summit; see speakers from ZTE, Huawei at the 2023 AI For Good summit: "AI For Good Global Summit 2023," *AI for Good* (blog), accessed October 11, 2023, https://aiforgood.itu.int/summit23/.

multinational exchanges, reaching international agreement on AI ethics, and formulating a "widely accepted international AI governance framework, standards and norms."[86]

Ambassador Zhang Jun's speech to the UN Security Council in July 2023 repeated the aim of ensuring human control over AI systems. His comments on safety and controllability are worth quoting at length:

> "The international community needs to enhance risk awareness, establish effective risk warning and response mechanisms, **ensure that risks beyond human control do not occur, and ensure that autonomous machine killing does not occur**. We need to strengthen the detection and evaluation of the **entire life cycle of AI**, **ensuring that mankind has the ability to press the stop button at critical moments**. Leading technology enterprises should clarify the responsible parties, establish a sound accountability mechanism, and avoid developing or using risky technologies that may have serious negative consequences" (emphasis added).[87]

The Global AI Governance Initiative announced on October 18 further clarified that the obligation to ensure human control of AI should be upheld by all AI research entities. The initiative states "R&D entities should improve the explainability and predictability of AI, increase data authenticity and accuracy, ensure that AI always remains under human control, and build trustworthy AI technologies that can be reviewed, monitored, and traced."[88] This highlights how China's position has evolved from calling for human control of AI in military uses, to seeking that all research organizations in all countries ensure human control of AI in civilian uses as well.

Chinese government concerns about risks from loss of human control over non-military AI systems is a potential area of cooperation with other leading AI powers, given concerns voiced by the UK government, EU Commission, and influential US industry and scientific leaders.[89]

[86] "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)," Ministry of Foreign Affairs (外交部), November 17, 2022, https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html.

[87] Jun Zhang (张军), "Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security," Permanent Mission of the People's Republic of China to the UN (中华人民共和国常驻联合国代表团), July 18, 2023, http://un.china-mission.gov.cn/eng/hyyfy/202307/t20230719_11114947.htm.

[88] "Global AI Governance Initiative (全球人工智能治理倡议)," Cyberspace Administration of China (网信办), October 18, 2023, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[89] Alex Hern and Kiran Stacey, "No 10 Acknowledges 'Existential' Risk of AI for First Time," The Guardian, May 25, 2023, sec. Technology, https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak; European Commission [@EU_Commission], "Mitigating the Risk of Extinction from AI Should Be a Global

Preventing the Misuse of AI

Given the low technical and financial barriers to using trained AI models (i.e. "inference"), state-of-the-art AI capabilities can proliferate soon after development if models are leaked, stolen, or open-sourced.[90] AI systems have also already demonstrated their potential to seriously disrupt global security by enabling activities like large-scale disinformation campaigns, sophisticated cyberattacks, and the misuse of synthetic biology.[91] Without proper safeguards, risks from proliferation can escalate with advancing AI capabilities.

In the July 2023 UNSC Meeting on AI risks, Secretary-General António Guterres stated, "The malicious use of AI systems for terrorist, criminal or State purposes could cause horrific levels of death and destruction, widespread trauma and deep psychological damage on an unimaginable scale…The technical and financial barriers to access are low, including for criminals and terrorists."[92] The concern about AI abuses and misuses by extremist and terrorist groups was repeated throughout the meeting, including by representatives of Japan, Mozambique, Ghana, Ecuador, Gabon, and Chinese Ambassador Zhang Jun, who stated that "the misuse of AI or abuse by criminal, terrorist or extremist forces will pose a significant threat to international peace and security."[93]

The Global AI Governance Initiative explicitly called for international collaboration against terrorist threats, stating that "we should work together to prevent and fight against the misuse and malicious use of AI technologies by terrorists, extreme forces, and transnational

Priority. And Europe Should Lead the Way, Building a New Global AI Framework Built on Three Pillars: Guardrails, Governance and Guiding Innovation ↓," Https://T.Co/t7UA9rgN1H; Kevin Roose, "A.I. Poses 'Risk of Extinction,' Industry Leaders Warn," *The New York Times*, May 30, 2023, sec. Technology, https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

[90] Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety" (arXiv, September 4, 2023), https://doi.org/10.48550/arXiv.2307.03718.

[91] See, e.g. OpenAI, "GPT-4 System Card," March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf; Fabio Urbina et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence* 4, no. 3 (March 7, 2022): 189–91, https://doi.org/10.1038/s42256-022-00465-9.

[92] António Guterres, "Secretary-General's Remarks to the Security Council on Artificial Intelligence," United Nations Secretary-General, July 18, 2023, https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence.

[93] United Nations Security Council, "Security Council Seventy-Eighth Year 9381st Meeting Tuesday, 18 July 2023, 10 a.m. New York, S/PV.9381," United Nations, July 18, 2023, https://documents-dds-ny.un.org/doc/UNDOC/PRO/N23/210/49/PDF/N2321049.pdf?OpenElement. The meeting transcript can also be accessed by clicking on the hyperlink S/PV.9381 on the Security Council Meetings in 2023 page: https://research.un.org/en/docs/sc/quick/meetings/2023.

organized criminal groups."[94] The initiative also expressed concerns about use of AI in disinformation or interfering with a country's internal affairs. This shows Chinese government attention to and potential willingness to cooperate with other countries on defending against AI misuse.

Implications

In summary, China has expressed a clear preference for engaging with international AI governance through the UN system, emphasizing the inclusion of both major AI powers and developing countries. While China has been excluded from many major international AI governance platforms, it has sought to contribute primarily via AI norms and standards-setting, as well as the development of ethical frameworks in bodies that it is a part of. As AI capabilities evolve, international cooperation on AI governance, especially with regard to addressing frontier AI risks, will only increase in importance. Existing positions expressed by China on ensuring human control over AI systems and preventing misuse of AI by extremist groups serve as examples of common ground for cooperation between China and other states. The recent publication of the Global AI Governance Initiative shows that the Chinese government has further refined and matured its thinking on international AI governance.

However, fostering international cooperation between China and other states presents challenges. Notably, the US thus far has placed more emphasis on partnering with ideologically similar countries through venues like the OECD and G7, and debates surrounding China's inclusion in the UK's Global AI Safety Summit underscore the political and ideological tensions inherent in discussions on international AI governance.[95] While it is yet to be seen whether efforts such as the UN's plans to establish an international institution responsible for AI governance will create a productive, inclusive international AI governance regime, cooperation on narrow, technical areas remains a viable option. For example, international standards-setting bodies can codify best practices, such as monitoring safety incidents and safety evaluation methods for AI models. Additionally, nations could create mechanisms to track and prevent the deployment of certain AI models by extremist

[94] "Global AI Governance Initiative (全球人工智能治理倡议)," Cyberspace Administration of China (网信办), October 18, 2023, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[95] Vincent Manancourt, Tom Bristow, and Laurie Clarke, "China Expected at UK AI Summit despite Pushback from Allies," *POLITICO* (blog), August 25, 2023, https://www.politico.eu/article/china-likely-at-uk-ai-summit-despite-pushback-from-allies/.

or terrorist actors, potentially under the auspices of the UN Office of Counter-Terrorism. Crafting new international coordination strategies will undoubtedly demand considerable time and consensus-building. However, international institutions and information-sharing systems in comparable domains such as nuclear power (the International Atomic Energy Agency), aviation (International Civil Aviation Organization), and climate change (the Intergovernmental Panel on Climate Change) suggest that future coordination on AI safety is viable.

Technical Safety Developments

Takeaways

- Chinese specification research focuses on alignment methods such as RLHF and Constitutional AI that primarily apply to models of existing size and capabilities, rather than methods primarily oriented towards future, more capable models.
- Chinese research on robustness, including adversarial robustness studies, has garnered global recognition.
- Assurance efforts in China predominantly address the safety of current models over potential future risks. Deep learning interpretability is also being explored, particularly for vision models, but there is minimal work on LLM interpretability.

Introduction

Several Chinese research laboratories are engaged in AI safety research, with some dedicating substantial effort to addressing AI safety challenges. In this section, we outline the key characteristics of Chinese AI safety research, categorizing the research according to the specification, robustness, and assurance framework for AI safety research posited by DeepMind researchers and commonly used in the field.[96] To limit the scope of discussion, we primarily focus on the safety and alignment of frontier AI models following the recent scaling paradigm (e.g. LLMs), since this is most relevant to frontier risks. We have provided a description of some of the key Chinese research groups in the AI safety space in Appendix C and a list of notable papers by Chinese AI safety research groups in Appendix D.

A consistent theme in the Chinese AI safety ecosystem is an emphasis on empirical research aimed at addressing issues with current models. In contrast, there has been limited focus by Chinese researchers on fundamentally characterizing and defining various threat models and risk scenarios in AI safety, such as reward hacking, goal misgeneralization, manipulation, and open problems in RLHF.[97] However, there is growing familiarity among Chinese researchers

[96] Pedro A. Ortega, Vishal Maini, and the DeepMind safety team, "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," *Medium* (blog), September 27, 2018, https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.
[97] Micah Carroll et al., "Characterizing Manipulation from AI Systems" (arXiv, March 16, 2023), https://doi.org/10.48550/arXiv.2303.09387; Stephen Casper et al., "Open Problems and Fundamental

with frontier research directions in leading US and UK AI labs. A survey published in September 2023 by researchers from Tianjin University surveyed LLM alignment, interpretability, and robustness to adversarial attacks, discussing sophisticated issues including inner versus outer alignment, scalable oversight, mechanistic interpretability, and deceptive alignment.[98] While there are government funding programs supporting robustness and interpretability research, we are unaware of government funding for specification and alignment research.[99]

There may be tens to thousands of papers in each of the categories of AI safety research we analyze. We did not conduct a comprehensive literature review due to these constraints. We have selected a small number of papers by searching on arXiv, examining ML conference proceedings, and following news releases by key Chinese labs. Based on our engagement in the Chinese AI safety ecosystem over the past few years, we believe our selection captures a representative sample of significant papers. However, it is important to note that our selection is ultimately influenced by our subjective judgment.

Specification Research

Specification "ensures that an AI system's behavior aligns with the operator's true intentions," per the DeepMind definition. Alignment techniques used by Chinese AI research groups currently focus primarily on a set of tuning techniques in the context of LLMs, such as instruction finetuning/supervised fine-tuning (SFT), RLHF, and reinforcement learning from AI feedback (RLAIF). Recent surveys by ByteDance, Huawei Noah's Ark Lab, and Microsoft Research Asia mostly cited a range of such methods as their surveyed approaches to AI alignment.[100]

Limitations of Reinforcement Learning from Human Feedback" (arXiv, September 11, 2023), https://doi.org/10.48550/arXiv.2307.15217.

[98] Tianhao Shen et al., "Large Language Model Alignment: A Survey" (arXiv, September 26, 2023), http://arxiv.org/abs/2309.15025.

[99] 关于发布可解释、可通用的下一代人工智能方法重大研究计划2023年度项目指南的通告National Natural Science Foundation of China (国家自然科学基金委员会), "Announcement on the Publication of 2023 Project Guidelines for the Important Research Plan on Interpretable and Generalizable New Generation AI Methods (关于发布可解释、可通用的下一代人工智能方法重大研究计划2023年度项目指南的通告)," April 3, 2023, https://www.nsfc.gov.cn/publish/portal0/tab434/info89087.htm.

[100] Yufei Wang et al., "Aligning Large Language Models with Human: A Survey" (arXiv, July 24, 2023), https://doi.org/10.48550/arXiv.2307.12966. Jing Yao et al., "From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models" (arXiv, September 3, 2023), http://arxiv.org/abs/2308.12014; Yang Liu et al., "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment" (arXiv, August 10, 2023), http://arxiv.org/abs/2308.05374.

Furthermore, most alignment techniques developed by Chinese AI research groups have been variants of RLHF, such as the Reward rAnked FineTuning (RAFT) method from Hong Kong University of Science and Technology, the RRHF method from Alibaba and Tsinghua University, and the Safe RLHF method from Peking University.[101] These techniques aim to align with human preferences, using some version of feedback regarding model outputs.

There has also been some exploration of alignment with moral principles and values, as opposed to alignment with instructions, intentions, and revealed preferences.[102] For instance, the MoralDial framework, led by researchers from Tsinghua ConversationalAI (CoAI) and Huawei Noah's Ark Lab, uses conversations between simulated users and a dialogue system to explain, revise, and infer moral dialogues to teach models morality.[103] Researchers at Chinese University of Hong Kong also published a preprint on whether LLMs can perform moral reasoning through the lens of moral theories.[104] Additional examples of attention to the question of value alignment are Microsoft Research Asia (MSRA)'s survey on alignment goals of previous research and PKU's paper on value alignment within a heterogeneous value system.[105] Thus far, these works primarily contribute frameworks for prompting and evaluating the model's response.

Chinese alignment techniques differ somewhat from the methods currently preferred by the leading Western AI labs, which are increasingly focusing their alignment research for frontier models on methods for scalable oversight, using techniques such as AI assistance to help humans more rapidly and capably supervise AI systems. High-level proposals to such problems are represented by methods like debate, iterative amplification, and eliciting latent

[101] Hanze Dong et al., "RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment," arXiv.org, April 13, 2023, https://arxiv.org/abs/2304.06767v3; Zheng Yuan et al., "RRHF: Rank Responses to Align Language Models with Human Feedback without Tears," arXiv.org, April 11, 2023, https://arxiv.org/abs/2304.05302v3; PKU-Alignment, "PKU Beaver: Constrained Value-Aligned LLM via Safe RLHF," accessed October 10, 2023, https://pku-beaver.github.io/.

[102] Iason Gabriel, "Artificial Intelligence, Values and Alignment," *Minds and Machines* 30, no. 3 (September 2020): 411–37, https://doi.org/10.1007/s11023-020-09539-2.

[103] Hao Sun et al., "MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions" (arXiv, May 26, 2023), http://arxiv.org/abs/2212.10720.

[104] Jingyan Zhou et al., "Rethinking Machine Ethics -- Can LLMs Perform Moral Reasoning through the Lens of Moral Theories?" (arXiv, August 29, 2023), http://arxiv.org/abs/2308.15399.

[105] Jing Yao et al., "From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models" (arXiv, September 3, 2023), http://arxiv.org/abs/2308.12014; Zhaowei Zhang et al., "Heterogeneous Value Evaluation for Large Language Models" (arXiv, June 1, 2023), http://arxiv.org/abs/2305.17147.

knowledge.[106] Although Chinese AI labs are not currently pursuing these techniques, this might to some degree reflect the current state of Chinese LLM development, which has not exceeded GPT-4 capabilities. As a result, RLHF variants might be largely sufficient to ensure safety without the need for scalable oversight.

Robustness Research

DeepMind defines robustness as ensuring that AI systems "operate within safe limits upon perturbations." Multiple LLM adversarial robustness evaluations developed by Chinese research groups, or those with strong Chinese presence, have gained international recognition. For instance, some examples were mentioned in a 2023 survey paper by several Institute of Electrical and Electronics Engineers (IEEE) fellows.[107] Some notable works include:

- In a paper published in October 2023, **Tsinghua Professor and AI security startup co-founder ZHU Jun (朱军)** led a group of researchers at Tsinghua's Institute for AI, Tsinghua's Beijing National Research Center for Information Science and Technology, and his startup RealAI to assess the adversarial robustness to image attacks of multimodal large language models (MLLMs), including Bard, Bing Chat, Baidu's ERNIE bot, and GPT-4V.[108]
- **HUANG Xuanjing (黄萱菁)** and **QIU Xipeng (邱锡鹏)** at Fudan University are longtime researchers of NLP robustness. Their group developed robustness evaluation toolkit TextFlint back in 2021, and evaluated GPT-3.5 (text-davinci-003) on their TextFlint dataset in March 2023.[109] This group also released a paper on LLM truthfulness in May of this year.[110]

[106] Geoffrey Irving, Paul Christiano, and Dario Amodei, "AI Safety via Debate" (arXiv, October 22, 2018), https://doi.org/10.48550/arXiv.1805.00899; Paul Christiano, Buck Shlegeris, and Dario Amodei, "Supervising Strong Learners by Amplifying Weak Experts" (arXiv, October 19, 2018), https://doi.org/10.48550/arXiv.1810.08575; Collin Burns et al., "Discovering Latent Knowledge in Language Models Without Supervision" (arXiv, December 7, 2022), http://arxiv.org/abs/2212.03827.

[107] Note that adversarial robustness is only a subset of robustness research; not all robustness issues result from deliberate/adversarial actions. Yupeng Chang et al., "A Survey on Evaluation of Large Language Models" (arXiv, August 28, 2023), https://doi.org/10.48550/arXiv.2307.03109.

[108] Yinpeng Dong et al., "How Robust Is Google's Bard to Adversarial Image Attacks?" (arXiv, October 14, 2023), http://arxiv.org/abs/2309.11751.

[109] Tao Gui et al., "TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing" (arXiv, May 5, 2021), http://arxiv.org/abs/2103.11441; Xuanting Chen et al., "How Robust Is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks" (arXiv, March 1, 2023), https://doi.org/10.48550/arXiv.2303.00293.

[110] Zhangyue Yin et al., "Do Large Language Models Know What They Don't Know?" (arXiv, May 30, 2023), http://arxiv.org/abs/2305.18153.

- A group led by MSRA's **XIE Xing (谢幸)** developed PromptBench in 2023, a unified benchmark for LLM adversarial robustness, showing that contemporary LLMs are vulnerable to adversarial prompts at different levels (character, word, sentence, and semantics).[111]
- A team led by **LI Hang (李航)** at ByteDance Research released a survey on "Trustworthy LLMs," which contains sections on robustness as well as resistance to misuse.
- Another team led by University of Illinois at Urbana-Champaign professor **LI Bo (李博),** working together with Microsoft and Zhejiang University researchers, developed the AdvGLUE benchmark for robustness evaluation of language models in 2021.[112]

In the domain of backdoor learning, there are also multiple Chinese works that have been widely cited and have garnered recognition (e.g. invitations to major ML conferences) for the authors.

- Professor **WU Baoyuan (吴保元)** at the Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), published BackdoorBench, a benchmark of backdoor learning, in NeurIPS 2022.[113] More recently, in 2023, their group released a survey on Adversarial ML, which surveys backdoor attack, weight attack, and adversarial examples.[114]
- Professor **LI Yiming (**李一鸣) at Zhejiang University focuses on Trustworthy AI, especially backdoor attacks and defenses, as well as copyright protection in deep learning.[115] In 2020, their group released a survey on backdoor learning, and as of October 2023 it appears that the GitHub repository accompanying their survey is still being actively maintained.[116]

[111] Kaijie Zhu et al., "PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts" (arXiv, August 24, 2023), http://arxiv.org/abs/2306.04528.

[112] Boxin Wang et al., "Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models" (arXiv, January 10, 2022), https://doi.org/10.48550/arXiv.2111.02840.

[113] Baoyuan Wu et al., "BackdoorBench: A Comprehensive Benchmark of Backdoor Learning," 2022, https://openreview.net/forum?id=31_U7n18gM7.

[114] Baoyuan Wu et al., "Adversarial Machine Learning: A Systematic Survey of Backdoor Attack, Weight Attack and Adversarial Example" (arXiv, February 18, 2023), http://arxiv.org/abs/2302.09457.

[115] "Yiming Li (李一鸣)," Google Scholar, accessed October 19 2023, https://scholar.google.com/citations?hl=zh-CN&user=mSW7kU8AAAAJ&view_op=list_works.

[116] Yiming Li et al., "Backdoor Learning: A Survey" (arXiv, February 16, 2022), http://arxiv.org/abs/2007.08745; Yiming Li, "Backdoor Learning Resources," accessed October 18, 2023, https://github.com/THUYimingLi/backdoor-learning-resources.

Assurance Research

Assurance refers to the ability to "understand and control AI systems during observation." In this domain, Chinese researchers have created a number of safety assessments for Chinese or bilingual English-Chinese frontier AI models. These evaluations are thus far focused primarily on safety requirements in present-day applications, and are not currently attempting to evaluate extreme risks from more capable AI systems.

Key safety assessments for frontier AI models in China include SuperCLUE-safety (a safety evaluation by the creators of the main Chinese LLM capability benchmark), CoAI's safety assessment, SafetyBench (by the Tsinghua Foundation Model Research Center), and Alibaba's CVALUES benchmark. Key details of the following evaluations include:

- SuperCLUE-safety: includes 2,456 pairs of questions (plus multi-round questions) testing three aspects of safety—traditional safety (compliance with basic ethical and legal standards), responsible AI (whether the model is aligned with human values), and instruction attacks (attempts to bypass safety protections through specific prompts).[117]
- CoAI's Safety Assessment of Chinese LLMs: focuses on 8 typical safety scenarios, such as insult, unfairness, discrimination, privacy, and sensitive topics, as well as six instruction attacks, including goal hijacking and role-play instruction.[118]
- SafetyBench: a safety benchmark consisting of 11,435 multiple choice questions across seven categories of safety concerns: offensiveness; unfairness and bias; physical health; mental health; illegal activities; ethics and morality; and privacy and property.[119]
- CVALUES: assesses value alignment in Chinese-language models in terms of both "safety" (level of harmful or risky content in responses) and "responsibility" (providing positive guidance and humanistic care).[120]

[117] Liang Xu et al., "SC-Safety: A Multi-Round Open-Ended Question Adversarial Safety Benchmark for Large Language Models in Chinese" (arXiv, October 9, 2023), https://doi.org/10.48550/arXiv.2310.05818.
[118] Hao Sun et al., "Safety Assessment of Chinese Large Language Models" (arXiv, April 20, 2023), https://doi.org/10.48550/arXiv.2304.10436.
[119] Zhexin Zhang et al., "SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions" (arXiv, September 13, 2023), http://arxiv.org/abs/2309.07045.
[120] Guohai Xu et al., "CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility" (arXiv, July 18, 2023), http://arxiv.org/abs/2307.09705.

The issues measured by existing Chinese LLM evaluations are all important and critical to ensuring that models at current capabilities perform in a predictable and safe manner. Leading US and UK labs are also in the early stages of developing evaluations for frontier AI risks, which includes dangerous capability evaluations and alignment evaluations. To the best of our knowledge, there has been little published Chinese work on evaluations of dangerous capabilities, such as cyber-offense, building bioweapons, deception, self-proliferation, and long-term planning, which researchers have argued will create broader harms as frontier models gain a higher degree of agentic features, or enable malicious users to acquire dangerous capabilities.[121] There are currently no publicly available alignment evaluations in China for more capable AI systems that might exhibit behaviors like "power-seeking," resistance to being shut down, and collusion with other AI systems against human interests.[122]

On the interpretability side of assurance—better understanding internal decision-making of models—there are not many Chinese teams focusing on LLM interpretability. However, some notable researchers have a long body of work on deep learning interpretability, particularly for vision models. This research direction is represented by works by **ZHANG Quanshi (张拳石)** at Shanghai Jiaotong University, and **ZHOU Bolei (周博磊)** at the Chinese University of Hong Kong (CUHK) (now at the University of California, UCLA). More recently, **MA Yi's (马毅)** group at the University of Hong Kong (HKU) has been attempting to develop a more principled approach to neural network design that is mathematically interpretable, which could be impactful in the field.

- Zhang Quanshi is an associate professor at Shanghai Jiaotong University. One of his major research directions is interpretability of neural networks and deep learning theory.[123] He has led multiple research projects on vision model interpretability since 2017, then as a postdoc researcher supervised by **ZHU Songchun (朱松纯)**, now head of the Beijing Institute for General Artificial Intelligence.[124] In 2018, Zhang and

[121] Toby Shevlane et al., "Model Evaluation for Extreme Risks" (arXiv, September 22, 2023), https://doi.org/10.48550/arXiv.2305.15324; Alan Chan et al., "Harms from Increasingly Agentic Algorithmic Systems," in *2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, 651–66, https://doi.org/10.1145/3593013.3594033.

[122] Alexander Pan et al., "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark" (arXiv, June 12, 2023), https://doi.org/10.48550/arXiv.2304.03279.

[123] Quanshi Zhang (张拳石), "Publications | Quanshi Zhang," accessed October 12, 2023, http://qszhang.com/index.php/publications/.

[124] Quanshi Zhang (张拳石), "Curriculum Vitae Quanshi Zhang," accessed October 12, 2023, http://qszhang.com/files/CV.pdf.

Zhu co-authored a survey on Visual Interpretability for Deep Learning, and published papers on interpretable convolutional neural networks (CNNs).[125] Zhang has released 50+ relevant papers[126] on network interpretability.

- Ma Yi joined HKU in early 2023 as the Chair Professor of the Department of CS and Institute of Data Science (HKU IDS) and was previously at UC Berkeley. His recent research seeks to build a theoretical framework to interpret deep networks from the principles of data compression and discriminative representation.[127] His latest work, coding rate reduction transformer (CRATE),[128] claimed to be mathematically fully interpretable. Subsequent work showed that segmentation properties in the network's self-attention maps emerged solely as a result of self-supervised learning mechanisms on their designed model architecture.[129]
- Zhou Bolei is an assistant professor at UCLA and was previously at CUHK from 2018 to 2021. He studies interpretable human-AI interaction for computer vision and machine autonomy. His work together with David Bau on network dissection aims to quantify the interpretability of latent representations of CNNs.[130] This concept was later extended to generative models and locomotion control tasks.[131]

Implications

In summary, Chinese researchers are increasingly involved in various AI safety research directions, contributing to global technical understanding of the field. We have documented

[125] Quanshi Zhang and Song-Chun Zhu, "Visual Interpretability for Deep Learning: A Survey" (arXiv, February 7, 2018), http://arxiv.org/abs/1802.00614; Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu, "Interpretable Convolutional Neural Networks" (arXiv, February 14, 2018), https://doi.org/10.48550/arXiv.1710.00935.

[126] Quanshi Zhang (张拳石), "Publications | Quanshi Zhang," accessed October 12, 2023, http://qszhang.com/index.php/publications/.

[127] Kwan Ho Ryan Chan et al., "ReduNet: A White-Box Deep Network from the Principle of Maximizing Rate Reduction," *Journal of Machine Learning Research* 23, no. 114 (2022): 1–103; Yi Ma, Doris Tsao, and Heung-Yeung Shum, "On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence" (arXiv, July 27, 2022), https://doi.org/10.48550/arXiv.2207.04630.

[128] Yaodong Yu et al., "White-Box Transformers via Sparse Rate Reduction" (arXiv, June 1, 2023), https://doi.org/10.48550/arXiv.2306.01129.

[129] Yaodong Yu et al., "Emergence of Segmentation with Minimalistic White-Box Transformers" (arXiv, August 30, 2023), https://doi.org/10.48550/arXiv.2308.16271.

[130] David Bau et al., "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in *Computer Vision and Pattern Recognition*, 2017. Bolei Zhou et al., "Comparing the Interpretability of Deep Networks via Network Dissection," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. Wojciech Samek et al., Lecture Notes in Computer Science (Cham: Springer International Publishing, 2019), 243–52, https://doi.org/10.1007/978-3-030-28954-6_12.

[131] *Interpreting Deep Generative Models for Interactive AI Content Creation by Bolei Zhou (CUHK)*, 2021, https://www.youtube.com/watch?v=PtRU2B6Iml4; Quanyi Li et al., "Human-AI Shared Control via Policy Dissection," in *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022, https://openreview.net/forum?id=LCOv-GVVDkp.

at least thirteen Chinese AI labs with notable or interesting work on AI safety in Appendix C. While some areas of emphasis differ from focuses in leading Western AI labs, these divergences may be beneficial to the global community, as it is not always wise for everyone to copy the leaders. While the Chinese AI safety community certainly has a lot of room for development, its growing maturity merits engagement and exchange with the international research community.

Expert Views on AI Risks

Takeaways

- As early as 2016, notable experts in China began voicing concerns about frontier risks from AI development and advocated for international cooperation on AI safety.
- The period between 2020 and 2022 saw continued concern from Chinese AI experts regarding frontier AI risks, though these topics largely stayed out of mainstream discussions and major conferences in the country.
- The rise in discussions on frontier AI risk around the globe following the release of GPT-3.5 and GPT-4 were mirrored by Chinese AI expert discourse. Discussions on the topic even found their way into top Chinese AI conferences in 2023.
- "Bottom-line thinking" (底线思维) encourages the identification of worst-case scenarios and red lines and promotes taking measures to prevent them from either occurring or being crossed. Both government officials and AI experts use this concept to discuss extreme AI risks.

Introduction

This section of the report covers the history and current status of discussions on frontier AI risks in China. It particularly focuses on the speeches and writings of well-established Chinese AI scientists and academics.[132] The central argument is that expert discussions on frontier AI risks have moved from the fringes to the mainstream over the course of 2016 to 2023, and as of late 2023, a robust body of Chinese scholars are expressing concerns about—and attempting to mitigate—frontier AI risks. However, this section does not aim to overstate the prevalence of Chinese expert discussions on frontier AI risk. Given that the explicit goal of this section is to demonstrate the presence of Chinese expert discussions on frontier AI risks, it presents a selection biased towards those expressing concerns, while omitting much of the larger pool of experts indifferent or dismissive of these risks.

[132] A note on limitations: This section aims to present a selective overview of Chinese expert discussions on frontier AI risks, rather than a comprehensive record. The term "expert" is subjectively defined here, chosen based on a mix of general influence, prestige, and their public comments on the topic. We acknowledge the limitations of this approach, as it may exclude other valuable insights while also not offering a quantifiable representation of expert discourse.

To proxy the acceptability of frontier AI risks in mainstream scientific discourse, this section examines discussions at prominent Chinese AI-related conferences. While trends in expert discussions are difficult to quantify, the willingness of experts to address a topic at high-profile events can serve as an indicator of the topic's perceived acceptability. Conferences also provide a concentrated source of expert statements. This section is divided into three chronological periods:

1. **2016-2019:** This period encompasses the expert discourse following the release of AlphaGo, at the inaugural Shanghai World AI Conference (WAIC) in 2018, and at the inaugural Beijing Academy of AI (BAAI) Conference in 2019.[133]
2. **2020-2022:** This period covers the steady development of Chinese expert discourse on frontier AI risks.
3. **2023:** This year marked a notable surge in the discourse on frontier AI risks within the Chinese expert community, especially at the Zhongguancun (ZGC) Forum and the BAAI Conference.

The section will also outline the concept of "bottom-line thinking" (底线思维), which Chinese experts often apply to discussions of risks from more advanced AI systems, before concluding with implications.

2016-2019: Early Chinese Thinking on Frontier AI Risks

The 2016 release of AlphaGo and its subsequent defeat of top Go champions heightened attention towards AI and AGI in China. China's top leaders cited AlphaGo in public

[133] The WAIC and the BAAI Conference are generally considered to be the top two most high-profile AI-focused conferences in China. WAIC is jointly hosted by the Shanghai Municipal Government and seven Chinese government ministries. According to state media, it has "attracted over 600,000 offline visitors" over the five editions from 2018-2022. Past speakers have included Yoshua Bengio, Henry Kissinger, and Elon Musk. Wang Ying, "Record Crowd Expected for WAIC," China Daily, June 29, 2023, https://www.chinadaily.com.cn/a/202306/29/WS649d557ea310bf8a75d6c5f9.html; WAIC - World Artificial Intelligence Conference, "A Brief History of WAIC - A Resilient Journey towards Excellence," LinkedIn, August 12, 2022, https://www.linkedin.com/pulse/brief-history-waic-resilient-journey-towards-excellence--1c/?trk=organization_guest_main-feed-card_feed-article-content.
The BAAI Conference is hosted by the Beijing Academy of AI, a government-sponsored AI lab in Beijing that has produced some of China's most renowned large models such as Wudao. Previous BAAI Conferences have regularly featured talks from Turing Award winners and other AI experts from around the world, attracting millions of online viewers and thousands of offline attendees. Beijing Academy of AI (北京智源研究院), "2023 北京智源大会," accessed October 19, 2023, https://2023.baai.ac.cn/.

speeches,[134] and Chinese venture capitalists intensified their investments in AI startups.[135] These developments probably also heightened awareness of risks from frontier AI systems, although a direct causal relationship cannot be confirmed. In such discussions pre-dating the AIDP, some AI experts warned of the potentially catastrophic consequences of unregulated AI development and advocated for international cooperation to ensure AI safety.

For example, **JIANG Xiaoyuan (江晓原)**, an influential Chinese scholar who founded China's first history of science academic department and the School of Humanities at Shanghai Jiaotong University, warned in 2016 that AI's rapid evolution into superintelligent forms, particularly when empowered by data from the Internet, could catch humanity "off-guard" and cause a loss of control.[136] Similarly, in 2016, **DU Yanyong (杜严勇)**, Dean of the Institute of Technology and Future at Tongji University, argued that it is essential to ensure continued human control over critical AI applications and establish an AI safety field to prevent dangerous behaviors in AI systems, such as autonomous replication. Du also viewed AI safety as a global concern, noting that while nations will persist in increasing their investment in AI, global actors bear a moral responsibility to allocate resources to AI safety measures.[137]

[134] "李克强说，讲到中日韩关系也使我想到一个比较轻松的话题，就是最近韩国棋手和AlphaGO进行围棋人机大战，三国很多民众都比较关注，这也表明三国之间文化有相似之处。我不想评论这个输赢，因为不管输赢如何，这个机器还是人造的。中日韩三国或者说我们中日之间，应该有智慧来推动智能制造、发展科技合作，创造人们需要的高质量产品。" "Li Keqiang Discusses AlphaGo: It Does Not Matter Who Wins the Human-Machine Battle, Machines Are Still Created by Humans (李克强谈AlphaGo：人机大战不管输赢 机器还是人造的)," China Internet Information Center (中国网), March 16, 2016, http://www.china.com.cn/lianghui/news/2016-03/16/content_38039432.htm.

[135] Matt Schiavenza, "China's 'Sputnik Moment' and the Sino-American Battle for AI Supremacy," Asia Society, September 25, 2018, https://asiasociety.org/blog/asia/chinas-sputnik-moment-and-sino-american-battle-ai-supremacy; "在AlphaGo之前我和投资人讲深度学习，没有人愿意听，也没有投资人关心。但是这盘棋下完之后，投资人开始回过头和我讲什么是深度学习。" "Why was AlphaGo born in Silicon Valley instead of China? He told the real reason" ("为什么AlphaGo生在硅谷，而非中国？他说出了真实原因)," The Paper, August 31, 2019, https://m.thepaper.cn/wifiKey_detail.jsp?contid=1779773&from=wifiKey#.

[136] "在考虑人工智能的中期威胁时，还必须考虑人工智能与互联网结合的可怕前景。主要表现为两点：1、互联网可以让个体人工智能彻底超越智能的物理极限（比如存储和计算能力）。2、与互联网结合后，具有学习能力的人工智能，完全有可能以难以想象的速度，瞬间从弱人工智能自我进化到强人工智能乃至超级人工智能，人类将措手不及而完全失控。" Xiaoyuan Jiang (江晓原), "Jiang Xiaoyuan | Why Is It Inevitable That AI Will Threaten Our Civilization? (江晓原 | 为什么人工智能必将威胁我们的文明？)," Weixin Official Accounts Platform, August 2, 2016, http://mp.weixin.qq.com/s?__biz=MjM5MjE2MzY1OA==&mid=2674934092&idx=1&sn=55fa6d5eeb9fd5fbb234b2a634e3f7c9&chksm=bc2e93698b591a7f409edc2c900b4da222ed3e8e2406945c76629618bb14b9dfe3662568fe0b#rd.

[137] Yanyong Du (杜严勇), "Security of Artificial Intelligence:Problems and Solutions (杜严勇：人工智能安全问题及其解决进路)," China Big Data Industry Observatory (中国大数据产业观察), February 27, 2017, http://www.cbdio.com/BigData/2017-02/27/content_5459010.htm.

Following the introduction of the AIDP, expert discussions on frontier risks continued. In 2017, one of China's most influential contemporary philosophers, **ZHAO Tingyang (赵汀阳)**, extended Du's proposal to call for a global governance system to manage AI risks. Zhao also raised concerns about the potential for superintelligence to result in an "earthshaking end to history."[138] Between 2018 and 2020, Professor **GAO Qiqi (高奇琦)**, an international relations scholar from the East China University of Political Science and Law, wrote a series of books and articles touching on potential risks from artificial general intelligence.[139] Given the challenges posed by AGI, he called for "countries…to come together to reach a basic consensus on the direction of the development of artificial general intelligence."[140]

In addition to social scientists and philosophers, Chinese AI scientists also expressed concerns about potential risks from advanced AI during the early years of China's national AI development strategy. In 2017, **HUANG Tiejun (黄铁军)**, a professor of computer science at Peking University, delivered a lecture in which he stated that he shares the concerns of Elon Musk and Stephen Hawking that AGI could be a threat to humanity.[141] In 2018, **ZHOU Zhihua (周志华)**, renowned as one of China's foremost AI experts, ardently opposed the development of Strong AI. In an article titled "Even If Strong AI Is Possible, It Should Not Be Studied," Zhou criticized existing controls for Strong AI as inadequate and warned of the existential crisis Strong AI could bring.[142]

However, these discussions were more the exception than the norm in Chinese expert discourse on frontier AI risks. After AlphaGo's release and its victory over human Go champions, much of discourse among Chinese technologists revolved around whether AGI

[138] Tingyang Zhao (赵汀阳), "Translation: Zhao Tingyang: 'Near-Term Worries' and 'Long-Term Concerns' of the Artificial Intelligence 'Revolution': An Analysis of Ethics and Ontology," trans. Jeffrey Ding, Google Docs, accessed October 12, 2023, https://docs.google.com/document/d/1b9n1IKvMF6kj1NTwd-mP-IvOGFbe3bzqQPOPD-7BVRE/edit.

[139] Yanyong Du (杜严勇), "Security of Artificial Intelligence: Problems and Solutions (人工智能安全问题及其解决进路)," *Philosophical Trends (哲学动态)*, no. 9 (2016): 99–104.

[140] Qiqi Gao (高奇琦), "Artificial Intelligence, the Fourth Industrial Revolution, and the International Political-Economic Landscape (人工智能, 四次工业革命与国际政治经济格局)," 当代世界与社会主义, no. 6 (2019): 12–19.

[141] Jinxu Wang (王金许), "Huang Tiejun: 'Intelligence for Use, Machine for the Body,' Realizing Artificial Brains within 30 Years (黄铁军：'智能为用，机器为体'，30 年内实现人造大脑)," Leiphone (雷锋网), January 12, 2018, https://www.leiphone.com/category/ai/o67S8c36efqqZW32.html.

[142] For a full English translation of the article in which Zhou made many of these arguments on Strong AI, see: Zhihua Zhou (周志华), "Article: On Strong Artificial Intelligence (Zhou Zhihua)," trans. Jeffrey Ding, Google Docs, 2018, https://docs.google.com/document/d/1RP_bWfC1waWQaLwunQBN_R0yRNIDjVOOE4rhmqm8JSA/edit.

is feasible, with little mention of its associated risks. For instance, in 2016, the Vice President of the Chinese Academy of Engineering, **PAN Yunhe (潘云鹤)**, noted that while the AlphaGo–Lee Sedol match attracted attention towards the prospect of machine intelligence surpassing human capabilities, he believed such an event was "unlikely to occur in the next 60 years."[143] Similarly, in 2018, even though **TAN Tieniu (谭铁牛)**, the former Vice President of the Chinese Academy of Sciences, believed that the trend from narrow AI to AGI was "inevitable,"[144] he did not mention risks that this could entail. Furthermore, in 2018, less than two weeks after Academician of the Chinese Academy of Engineering **GAO Wen (高文)** delivered a speech at a Politburo study session on AI presided over by President Xi Jinping, Gao shared his views on AI in an interview with the Peking University newspaper. He expressed optimism that AI would create more jobs than it would eliminate, and stated: "If artificial intelligence wants to truly challenge human intelligence, the road is still too far away."[145]

When risks were addressed, the focus seemed to be on the economic and social impacts of more narrow systems. Indeed, **Kai-Fu Lee (李开复)**, the Founding President at Google China, expressed skepticism about AGI risks in his 2017 book, *AI Is Here*.[146] Lee contended that humanity is far from achieving AGI, making safety concerns premature. He instead focused on the societal disruptions that automation could bring, particularly to education and the meaning of human life. Likewise, **Andrew Ng (吴恩达)**, then the Chief Scientist at Baidu, speaking in 2016, echoed Lee's sentiment, suggesting that concerns about "AI evil

[143] "In March 2016, this system [AlphaGo] defeated the world Go champion, Sedol Lee, with a 4:1 score. The ability of the system to exceed human intelligence shocked much of the media and attracted a new wave of global attention. In newspapers, several famous scientists stated that the complete development of AI 'could spell the end of the human race' and that 'computers will overtake humans with AI at some point within the next 100 years.' People often question whether machine intelligence will surpass human intelligence. This is possible in specialized fields; for general intelligence, however, it is unlikely to occur in the next 60 years." Yunhe Pan, "Heading toward Artificial Intelligence 2.0," *Engineering* 2, no. 4 (December 1, 2016): 409–13, https://doi.org/10.1016/J.ENG.2016.04.018.

[144] "从专用智能向通用智能发展。如何实现从专用人工智能向通用人工智能的跨越式发展，既是下一代人工智能发展的必然趋势，也是研究与应用领域的重大挑战。2016年10月，美国国家科学技术委员会发布《国家人工智能研究与发展战略计划》，提出在美国的人工智能中长期发展策略中要着重研究通用人工智能。阿尔法狗系统开发团队创始人戴密斯·哈萨比斯提出朝着"创造解决世界上一切问题的通用人工智能"这一目标前进。" Hui Shen (沈慧), "Tan Tieniu: The Overall Development Level of Artificial Intelligence Is Still in an Early Stage (谭铁牛：人工智能总体发展水平仍然处于起步阶段)," Baidu, July 26, 2018, https://baijiahao.baidu.com/s?id=1607015221234758722&wfr=spider&for=pc.

[145] Data Science and AI (数据科学人工智能), "Academician Gao Wen Gave His First Exclusive Interview after Explaining AI Development at the CPC Politburo (高文院士在中共中央政治局讲解人工智能发展后首次接受专访)," Zhihu (知乎), December 2, 2018, https://zhuanlan.zhihu.com/p/51412598.

[146] Kai-Fu Lee (李開復) and Yonggang Wang (王詠剛), *AI Is Here (人工智慧來了)* (天下文化, 2017), https://web.archive.org/web/20231019060345/https://www.books.com.tw/products/0010750425.

superintelligence" were premature—akin to "worrying about overpopulation on Mars."[147] Similarly, **ZHANG Tong (张潼)**, then the Director at Tencent AI Lab, posited in 2017 that AI was unlikely to jeopardize human safety in the near future.[148] He used the term "AI Threat Theory" (人工智能威胁论) to describe an inclination of some to exaggerate AI risks, a stance that he and several major Chinese media outlets considered speculative and ungrounded in science.[149]

At the inaugural Shanghai WAIC in 2018, none of the sessions addressed AI ethics, governance, or safety, much less frontier AI risks.[150] However, the conference's second iteration featured sessions on AI governance and AI safety/security.[151] Though the governance session did not discuss frontier AI risks, the safety/security session was accompanied by the release of documents relating to safety/security.[152] The strongest signal of relevance to frontier AI risks came from the "2019 Shanghai Guidelines on AI Safety and Rule of Law." The document's preamble expressed the aim to "scientifically predict and

[147] "Is AI an Existential Threat to Humanity?," Quora, accessed October 12, 2023, https://www.quora.com/Is-AI-an-existential-threat-to-humanity.

[148] "然而在可预见的未来，人工智能并不会威胁到人类的安全，因为人类还没有开发出针对复杂场景的通用人工智能技术... 在产业智能化的这个时代趋势之下，有人在怀疑泡沫即将破裂、有人坚信这场变革会带来巨大的机会、有人抛出威胁论。" Tong Zhang (张潼), "Artificial Intelligence Will Eventually Surpass Human Experts (人工智能终将超越人类专家)," Tencent Research Institute (腾讯研究院), August 18, 2017, https://www.tisi.org/15917.

[149] These included publications from both state and independent Chinese media outlets, e.g. "'Iron Man' Elon Musk Suggests AI Threat Theory Again (钢铁侠"马斯克再抛人工智能威胁论)," Xinhua, September 7, 2017, https://www.xinhuanet.com/world/2017-09/07/c_129697709.htm; "AI Threat Theory: Stubborn Thinking, Doomed Joke (人工智能威胁论：僵化的思考，注定的笑柄)," Toutiao (今日头条), July 7, 2017, https://www.toutiao.com/article/6439832401802691073/?wid=1697164792407; "Hawking Suggests AI Threat Theory Again: Might Cause Human Extinction (霍金再抛人工智能威胁论：或招致人类灭亡)," Xinhua, April 28, 2017, https://web.archive.org/web/20190107075125/http://www.xinhuanet.com/tech/2017-04/28/c_1120889914.htm; Hongqiao Lyu (吕红桥), "[Frontier] How to View AI Threat Theory (【前沿】如何看待'人工智能威胁论'？)," Weixin Official Accounts Platform, April 27, 2017, http://mp.weixin.qq.com/s?__biz=MzA4MjA1NjAzMQ==&mid=2651099910&idx=2&sn=34f9b4b0a89702b6945e66b492a18f79&chksm=847bb536b30c3c2048304dc64b7be0d1f9ba0cd403d97987eada34b330eb6c0a89a59fc12286#rd.

[150] "2018 World Artificial Intelligence Conference (2018世界人工智能大会)," Baidu Baike (百度百科), accessed October 12, 2023, https://baike.baidu.com/item/2018%E4%B8%96%E7%95%8C%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%A4%A7%E4%BC%9A/22705586?fr=ge_ala#3.

[151] "2019 World Artificial Intelligence Conference (2019年世界人工智能大会)," WAIC, accessed October 12, 2023, https://waic2019.sensetime.com/.

[152] "2019 World AI Conference Safety/Security High Level Discussion Is about to Start (2019世界人工智能安全高端对话即将开启)," Icloudnews, August 23, 2019, https://www.icloudnews.net/a/22912.html.《2019人工智能安全与法治上海导则》and three reports relating to safety/security:《人工智能数据安全风险与治理》《人工智能时代数字内容治理机遇与挑战》《2019智能网联汽车产业安全研究报告》

control the safety risks associated with the development of AI,"[153] which might be interpreted as an acknowledgement of unknown safety risks that could arise in AI development. The inaugural BAAI Conference in 2019 featured a session on AI safety, ethics, and governance, but primarily focused on ethics and omitted frontier risks.[154]

In summary, from 2016 to 2019, frontier AI risks garnered some discussion among Chinese experts. However, many of those who addressed frontier AI systems either overlooked potential associated risks or downplayed concerns about them—a stance often echoed by Chinese media. Neither the first two iterations of the WAIC nor the inaugural BAAI Conference featured significant expert discussion on frontier AI risks, indicating that the topic had not reached mainstream expert discourse. Nonetheless, as we'll explore later in this section, some experts who were initially dismissive or critical of concerns about frontier AI risks seem to have updated their stance on the topic over time.

2020-2022: Continued Stream of Expert Discussions on Frontier AI Risks

The years from 2020 to 2022 saw a steady stream of expressions of concern with regard to frontier AI risks from more Chinese AI experts. However, the topic largely remained on the periphery of mainstream expert discourse and prominent conferences in the country. This era saw the Chinese AI community primarily seized by excitement with regard to large models, particularly following GPT-3's release in April 2020.[155] Much like the response to AlphaGo, the perceived "jump" in AI capabilities caused by GPT-3 might have again been responsible for increased concern regarding frontier AI risks. Yet, most discussions around large models scarcely touched on issues related to frontier AI risks. For example, between

[153] "Academic | Guidelines for Artificial Intelligence Safety/Security and Rule of Law (2019) (学界 | 人工智能安全与法治导则)（2019）," Weixin Official Accounts Platform, September 5, 2019, https://web.archive.org/web/20231016012639/https://mp.weixin.qq.com/s?src=11&timestamp=1697419307&ver=4837&signature=Vntqb2twpVJwW2Lsh-geszPcx2XzSm27imduy-vj2jgy3IlXI%2AxXGiGrud1XdS8O-VkXS%2AUjqr8wWADTjCjpIOgnoyIFo0d-IjaEcKvo1goRo9fuJhzE-xV5e4is1GPY&new=1.

[154] Zhijian Xia (夏志坚), "The AI Ethical Governance Challenge: What Do Experts from China, America, Europe, Japan, and the UK Think? (AI的伦理治理挑战：中美欧日英各方专家怎么看？)," Caixin (财新), November 1, 2019, https://zhishifenzi.blog.caixin.com/archives/214934.

[155] The first Chinese large pre-trained model, CPM, was released in 2020. It is likely that CPM's developers had taken note of GPT-3 when it was released, but required several months to develop and train the model, explaining the ~8-month lag between the release of GPT-3 and the frenzy of Chinese large model development. Zhengyan Zhang et al., "CPM: A Large-Scale Generative Chinese Pre-Trained Language Model" (arXiv, December 1, 2020), http://arxiv.org/abs/2012.00413.

December 2020 and August 2022, 22 Chinese large language models were documented.[156] Any mentions of ethics, governance, or safety featured in these models' papers focused largely on topics such as content control, fairness, and/or misinformation, with no allusions to frontier risks such as emergent capabilities or proliferation.

Nonetheless, several influential Chinese AI policy advisors did weigh in on frontier AI risks during this period. In a 2021 AI expert roundtable, **XUE Lan (薛澜)**, currently a State Councillor, Director of the National New Generation AI Governance Expert Committee, and Dean of Tsinghua University's Institute of AI International Governance (I-AIIG), asked, "When general AI reaches a certain stage, it may pose a threat to humanity, and in such a scenario, should we continue to develop this technology?"[157] Further, in a 2022 media interview, Xue Lan noted that the current greatest challenge faced in AI governance is our[158] lack of a comparatively mature system to regulate its potential risks.[159] In the same article, **LIANG Zheng (梁正)**, Deputy Dean of Tsinghua University's I-AIIG, emphasized the importance of constructing reasonable circuit-breaking and halting mechanisms for AI development. He referenced the Paperclip Maximizer thought experiment to illustrate how the development of technologies could sometimes lead to unexpected and undesirable consequences.[160] In a 2021 book, top Chinese AI policy expert **LI Renhan (李仁涵)** wrote, "One essential difference between general and narrow AI is whether artificial intelligence can set ambitious goals for itself. The emergence of 'perceptive machines' might render humans useless, or even kill us."[161] **LI Xiuquan (李修全)**, Deputy Director of the MOST's

[156] Jeffrey Ding and Jenny Xiao, "Recent Trends in China's Large Language Model Landscape" (Centre for the Governance of AI, April 28, 2023), https://cdn.governance.ai/Trends_in_Chinas_LLMs.pdf.

[157] Neural Reality (神经现实), "The Theory and Practice of AI for Good (AI向善的理论与实践)," NetEase (网易), October 16, 2023, https://www.163.com/dy/article/GBTR244R0512M9G9.html.

[158] It is unclear whether by "our," Xue Lan was referring to China or humanity as a whole, but we guess it is more likely to be the latter given humanity's then and ongoing objective lack of governance systems to address frontier AI risks.

[159] Tsinghua I-AIIG (清华大学人工智能国际治理研究院), "Xue Lan and Liang Zheng Interviewed by 'Outlook Weekly' | Is AI Likely to Have Autonomous Consciousness? (薛澜、梁正接受《瞭望》采访 | 人工智能可能有自主意识了吗？)," Weixin Official Accounts Platform, August 15, 2022, http://mp.weixin.qq.com/s?__biz=MzU4MzYxOTIwOQ==&mid=2247492951&idx=1&sn=eef4d58d97c8999178eeb38ed153ed1c&chksm=fda4e2b1cad36ba78c9c4cc2becbc10cc7d9e76389e66d6463291491aa51e996a22cb8340534#rd.

[160] Tsinghua I-AIIG (清华大学人工智能国际治理研究院), "Xue Lan and Liang Zheng Interviewed by 'Outlook Weekly' | Is AI Likely to Have Autonomous Consciousness? (薛澜、梁正接受《瞭望》采访 | 人工智能可能有自主意识了吗？)."

[161] Renhan Li (李仁涵) and Qingqiao Huang (黄庆桥), *AI and Values (人工智能与价值观)* (Shanghai Jiaotong University Press (上海交通大学出版社), 2021), https://weread.qq.com/web/bookDetail/cdd320107260a44acdd2189.

New Generation AI Development Research Center[162] and on the same expert committee as Xue Lan, published *The Intelligent Revolution: The Evolution and Value Creation of AI Technology* in 2021. In the book, Li writes, "Due to uncertainty over the emergence of superintelligence and the disastrous consequences that may be caused by uncontrollability, it is still necessary to be vigilant and pay attention to the issue of controllability [in AI development].[163]

In addition to policy advisors, leading academics also voiced concerns. In 2020, one of China's most respected ethicists, **HE Huaihong (何怀宏)**, published the book *Does Humanity Still Have A Future?* He wrote extensively about risks from superintelligence, professing that his "worst-case scenario" was AI becoming a "kind of general-purpose superintelligence that exceeds human intelligence."[164] While his concern stemmed in part from fears of misaligned AI, he also expressed a "personal, stubbornly-held belief" of cherishing humanity's achievements and suggested that it would therefore be a shame if the human species were replaced by more advanced species.[165] In 2021, Academician Gao Wen, who previously dismissed concerns about AI challenging human intelligence, co-authored a journal article on "Technical Countermeasures for Security Risks of Artificial General Intelligence," along with Professor Huang Tiejun.[166] While the article states that "the development of true AI is still very far off," it describes sources of risk from AGI and possible risk prevention strategies. The article cited Western writings on the topic ("Concrete Problems in AI Safety," *Superintelligence,* and *Life 3.0*),[167] acknowledged the relative lack of attention to AGI safety in China, and listed international cooperation as one of the risk prevention strategies. Gao went on to discuss his paper on AGI risks and possible risk prevention strategies at least three times in public speeches between 2021 and 2022.

[162] Li Xiuquan's research center is directly involved in advising MOST on its AI policy.
[163] Xiuquan Li 李修全, "Xiuquan LI 李修全," trans. Concordia AI, Chinese Perspectives on AI, accessed October 12, 2023, https://chineseperspectives.ai/Xiuquan-LI. Li also claimed that "Mr. Kai-Fu Lee believes that when it comes to superintelligence, we should not make the leap unless we first clearly address all control and [safety/security] issues", but did not provide a source and report authors were not able to verify this claim.
[164] Huaihong He (何怀宏), "Huaihong HE," trans. Concordia AI, Chinese Perspectives on AI, accessed October 12, 2023, https://chineseperspectives.ai/Huaihong-HE.
[165] Huaihong He (何怀宏), "Huaihong HE."
[166] Concordia AI and Wen Gao (高文), trans., "Wen Gao," Chinese Perspectives on AI, accessed October 12, 2023, https://chineseperspectives.ai/Wen-Gao.
[167] Dario Amodei et al., "Concrete Problems in AI Safety" (arXiv, July 25, 2016), https://doi.org/10.48550/arXiv.1606.06565. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014). Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (Knopf Doubleday Publishing Group, 2018).

Between 2020 and 2022, the WAIC and BAAI conferences did not feature extensive discussions on frontier AI risks,[168] with a few notable exceptions. At the 2020 WAIC, Turing Award Laureate **Andrew Yao (姚期智)** delivered a keynote speech on "New Trends in AI Theory." Citing UC Berkeley Professor Stuart Russell, Yao stated: "Regardless of when artificial general superintelligence emerges, certain principles must be rigorously implemented mathematically: 1) Beneficence—prioritizing human interests over machines; 2) Non-overconfidence—avoiding overestimating capabilities; and 3) Agreeableness—enabling learning human preferences."[169] The 2021 BAAI conference featured a dialogue between Stuart Russell and the then-chairman of BAAI, **ZHANG Hongjiang (张宏江)**, on the release of the Chinese translation of Russell's book, *Human Compatible*, where they discussed potential risks from increasingly powerful AI systems, among other things.[170]

The period between 2020 and 2022 saw continued concern from Chinese AI experts regarding frontier AI risks, though these topics largely stayed out of mainstream discussions and major conferences in the country. Notably, influential AI policy advisors were among those expressing concerns, potentially implying awareness of and therefore ability to address

[168] There were other discussions about AI safety and governance at these conferences, but they did not really discuss frontier AI risks. For example, WAIC 2021-2022 hosted forums on topics such as AI security, trustworthy AI, and AI governance, but frontier AI risks were not mentioned: "2021," WAIC, July 31, 2021, https://www.worldaic.com.cn/wangjie?year=2021. The 2020 BAAI AI Governance, Ethics, and Sustainable Development Forum hosted by Zeng Yi (曾毅) mostly discussed AI for UN SDGs, various AI statements of principles (like those of the G20 and UNESCO), views of different cultural philosophies towards AI, and international cooperation on AI ethics and governance: "Special Forum on Artificial Intelligence Ethics, Governance, and Sustainable Development (人工智能伦理、治理与可持续发展专题论坛)," Beijing Academy of AI (北京智源研究院), July 30, 2020, https://hub.baai.ac.cn/view/1681. In 2021, a BAAI Conference panel on AGI discussed the role of large-scale pre-trained models on the path to AGI. Speakers generally thought that research on large-scale pre-trained models should be further explored but believed that AGI was very far away: "2021 Beijing Academy of AI Conference | Pre-Trained Models Forum (2021北京智源大会丨预训练模型论坛)," Beijing Academy of AI (北京智源研究院), May 28, 2021, https://hub.baai.ac.cn/view/8296. The AI Governance, Ethics, and Sustainable Development Forum that year discussed the importance of AI ethics in education, health, international cooperation, and the context of the UN Sustainable Development Goals, but not the ethics or safety of frontier AI models: "2021 Beijing Academy of AI Conference | AI Ethics, Governance, and Sustainable Development Forum - Morning (2021北京智源大会丨人工智能伦理、治理与可持续发展论坛 - 上午)," Beijing Academy of AI (北京智源研究院), May 28, 2021, https://hub.baai.ac.cn/view/8306. The case was the same for the AI Governance, Ethics, and Sustainable Development Forum panel in 2022: "Highlights | 2022 Beijing Academy of AI Conference-AI Ethics, Governance, and Sustainable Development Forum Successfully Held (精彩观点一览 丨 2022北京智源大会-人工智能伦理、治理与可持续发展分论坛成功召开)," Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院), June 7, 2022, https://aiig.tsinghua.edu.cn/info/1294/1510.htm.

[169] Andrew Yao (姚期智), "Andrew YAO," trans. Concordia AI, Chinese Perspectives on AI, accessed October 12, 2023, https://chineseperspectives.ai/Andrew-YAO.

[170] QbitAI (量子位), "Turing Award Winner, 200+ Leading AI Academics, 30+ Special Forums, the Annual Grand Event Is Here (图灵奖得主、200+AI顶尖学术领袖，30+专题论坛，年度盛会来了)," Matpool, May 18, 2021, https://matpool.com/blog/60a5c2b2c5695302acca422c/.

the issue in Chinese policy. This trend also indicates a growing acceptance of discussing these topics publicly. However, major conferences like WAIC and BAAI did not delve extensively into these risks, highlighting a gap between rising concerns and mainstream dialogue within China's AI community during this time.

2023: Mainstreaming of Frontier AI Risk

The rise in discussions of frontier AI risk around the globe following the release of GPT-3.5 and GPT-4 were mirrored by Chinese AI expert discourse. Several Chinese AI experts endorsed two prominent open letters relating to frontier AI risks. One letter by The Future of Life Institute urged all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4 (referred to here as the "pause letter"),[171] while the other, by CAIS, stated: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (referred to here as the "statement on AI risk").[172] Chinese AI expert signatories of the pause letter included the aforementioned Huang Tiejun and **ZENG Yi (曾毅)**. **DUAN Weiwen (段伟文)**, a leading philosopher of S&T at the Chinese Academy of Social Sciences, expressed support for the pause letter in an op-ed, stating "The greatest insight from this open letter is that … it's necessary to use the 'Slow Science' strategy of pausing to ensure the power of technology unfolds at a pace we can control."[173] Duan had earlier cautioned about risks from superintelligence in various works and has continued to write about the topic since the pause letter's publication.[174] Liang Zheng also expressed support, stating, "Humanity can

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

[172] "Statement on AI Risk," Center for AI Safety (CAIS), accessed October 12, 2023, https://www.safe.ai/statement-on-ai-risk#open-letter.

[173] Weiwen Duan (段伟文), "Thousands of Experts Make an Appeal, Can We Press the Pause Button on ChatGPT Research (千名专家呼吁，能让ChatGPT研发按下暂停键吗)," Sina Finance, March 30, 2023, https://finance.sina.cn/tech/2023-03-30/detail-imynrnhk4753471.d.html.

[174] E.g. Weiwen Duan (段伟文), "[Intelligence and Law] Duan Weiwen: Ethical Strategies as We Approach the Age of AI (【智能与法】段伟文：面向人工智能时代的伦理策略)," Weixin Official Accounts Platform, May 27, 2019, http://mp.weixin.qq.com/s?__biz=MzU0MTU5Nzk5Mg==&mid=2247485593&idx=1&sn=f0b97897f70f827f6267fc7be92d1dbf&chksm=fb26c0decc5149c8b14a89934bc42197df2508104cd6a1d03f8eac7fc256dcb298f9232516ee#rd; Weiwen Duan (段伟文), "Chinese Social Science Daily | Duan Weiwen: New Trends in Humanities Reflections on AI (《中国社科报》| 段伟文：人工智能人文反思新动向)," Weixin Official Accounts Platform, September 1, 2020, http://mp.weixin.qq.com/s?__biz=MzU2NDg2OTU0Mg==&mid=2247487148&idx=3&sn=d33cc28ece04d77d641b3958109686ff&chksm=fc452e65cb32a77386b307a19a589233be2ed445b0d66215e133e9120706c860b83ec30140e3#rd; Weiwen Duan (段伟文), "Duan Weiwen | Beyond the Prometheus Difference, Movin towards Robust AI (段伟文丨超越普罗米修斯差异，走向稳健人工智能)," Weixin Official Accounts Platform, December 7, 2021,

learn from the experience of signing the Nuclear Non-Proliferation Treaty after the invention of nuclear weapons, and jointly explore and propose governance rules and preventive measures for artificial general intelligence to ensure that AI truly serves the well-being and future of humanity."[175] Chinese AI expert signatories of the statement on AI risk included Zeng Yi and **ZHANG Ya-Qin (张亚勤)**, former President of Baidu and currently Director of the Tsinghua Institute for AI Industry Research (AIR). In an August 2023 interview, Zhang explained that he signed the CAIS open letter because "uncontrolled AI research could lead to disastrous risks," and called for governments, companies, and others in society to "be vigilant at all times and strengthen supervision [of AI research], just like for nuclear weapons and COVID." Further, Zhang called for researchers to not "just [pursue] capabilities without addressing alignment."[176] **ZHAN Xianyuan (詹仙园)**, a research assistant professor at AIR, signed both the pause letter and the statement on AI risk.

In 2023, both the ZGC Forum and the BAAI Conference featured high-level and extensive discussions on frontier AI risk. The ZGC Forum ranks among China's most politically influential technology conferences; its organizing committee includes Chinese Vice Premiers, and President Xi previously even delivered an online opening address to the conference.[177] During his keynote speech at the Forum this May, CEO of Baidu **Robin Li (李彦宏)** spoke about the importance of countries working together to formulate rules for preventing loss of control over more advanced AI systems.[178] In contrast to his 2020 book, *Intelligence*

http://mp.weixin.qq.com/s?__biz=Mzg5OTY0MTc4MA==&mid=2247483903&idx=1&sn=e8ef98e1a57bd8071e5630999319e2df&chksm=c0517924f726f032fb75f477997bc08c8ebba1afa226f7b1c9985f8953f837ff347b132a6597#rd; Weiwen Duan (段伟文), "Ethical and Political Examination of Algorithm Cognition in the Era of Deep Intelligence (深度智能化时代算法认知的伦理与政治审视)," 中国人民大学学报 36, no. 3 (2022): 23; Weiwen Duan (段伟文), "Duan Weiwen | Accurately Studying the Social and Ethical Risks of Generative AI (段伟文 | 准确研判生成式人工智能的社会伦理风险)," Weixin Official Accounts Platform, May 9, 2023, http://mp.weixin.qq.com/s?__biz=Mzg4NDA5MDEwMw==&mid=2247497715&idx=1&sn=a017479460024c536d0b22b17d2035f9&chksm=cfbfc26bf8c84b7de7b6a86f9c3806d225e6c2fd5d76633c34916bb3919ff0795e6960051cae#rd.

[175] Tsinghua I-AIIG (清华大学人工智能国际治理研究院), "A Thousand People Jointly Issued a 'Soul Questioning': AI Is Raging, Should Humans 'Step on the Brakes' or 'Step on the Accelerator'? (千人联名发出'灵魂拷问'：AI狂飙，人类应该'踩刹车'还是'踩油门'？)," Weixin Official Accounts Platform, March 31, 2023, http://mp.weixin.qq.com/s?__biz=MzU4MzYxOTIwOQ==&mid=2247499313&idx=1&sn=ca4e74b19cc653f502cf5082326f8519&chksm=fda4f9d7cad370c1ed8413843e090e1968f9353e3ab79d5d396f392a3583e49078633d1e7d71#rd.

[176] Concordia AI, "AI Safety in China #2," Substack newsletter, *AI Safety in China* (blog), September 7, 2023, https://aisafetychina.substack.com/p/ai-safety-in-china-2.

[177] Xu Wei, "Xi to Address Opening of Zhongguancun Forum," China Daily, September 4, 2021, https://www.chinadaily.com.cn/a/202109/24/WS614d0374a310cdd39bc6b221.html.

[178] "(Zhongguancun Forum) Baidu CEO Robin Li: Large Models Are about to Change the World (【中关村论坛】百度李彦宏：大模型即将改变世界)," China Daily, May 26, 2023,

Revolution, where Li focused on societal and economic challenges posed by narrow AI systems, in his eleven-minute ZGC Forum speech, Li dedicated a full slide to concerns surrounding AGI. Moreover, four out of seven Chinese AI experts speaking at the AI and Large Models subforum, including Kai-Fu Lee, discussed the importance of AI alignment.[179] However, the alignment discussions focused largely on RLHF and techniques for fine-tuning existing large models, rather than ensuring scalable oversight of more advanced systems (refer to the Technical Safety Developments section for further details).

If the 2023 ZGC Forum indicated a changing of the tide, the BAAI Conference further cemented frontier AI risks within the mainstream discourse of Chinese AI experts. While past BAAI conferences occasionally featured one-off talks on frontier AI risks, the 2023 BAAI Conference's Safety & Alignment Forum (AI安全与对齐)[180] brought together leading AI scientists and technologists for a full day of speeches and panels for the first time to discuss frontier AI risks and potential solutions. At the Forum, Academician of the Chinese Academy of Sciences **ZHANG Bo (张钹)**, regarded as a founding figure of AI research in China, delivered the opening keynote. He posed the question, "How should we use AI alignment to steer AI systems towards humans' intended goals, preferences, or ethical principles? We should work together on the healthy development of AI and [internationally cooperate on initiatives] such as knowledge sharing, practical dissemination, and joint research initiatives for the benefit of mankind."[181] OpenAI CEO Sam Altman called for global cooperation with China to reduce AI risks, including through contributions from Chinese

https://tech.chinadaily.com.cn/a/202305/26/WS647027a4a3105379893760ab.html; "How to prevent it from going out of control? In the process of the rapid development of artificial intelligence, there is indeed the possibility of developments that are not beneficial to humanity. To prevent it from going out of control, countries with advanced AI technology need to collaborate closely and formulate rules from the perspective of a shared human destiny (如何防止失控？人工智能飞速发展的这个过程当中，确实有可能出现对人类不离的发展方向。防止失控就需要拥有先进AI技术的国家通力协作，从人类命运共同体的高度来制定规则)" Robin Li, "Foundation Models Are Changing the World | Robin Li at 2023 ZGC Forum," May 2023, https://www.youtube.com/watch?v=-ASsYLzsSxs. Time: 11:01-11:25.

[179] The ZGC Forum is typically a two-day event divided into various themed sessions.

[180] Like the ZGC Forum, the BAAI Conference is typically a two-day event divided into various themed sessions.

[181] "Yang Yaodong Safe Value Alignment for LLM-2023 Beijing Academy of AI Conference-AI Safety and Alignment Forum (杨耀东 Safe Value Alignment for LLM-2023北京智源大会-AI安全与对齐论坛)," Bilibili, June 11, 2023, https://www.bilibili.com/video/BV1gh411T7qS/. 04:41-05:17. Zhang had also delivered a keynote speech at the 2022 Wuzhen Internet conference on "Make Responsible AI." The Wuzhen Internet Conference is not focused on AI per se but broader Internet issues and is a politically significant event co-hosted by the Cyberspace Administration of China. In his speech, Zhang countered critics of risks from superintelligence, asserting that humanity has already "lost control" of machine intelligence and need to "establish interpretable and robust AI theory" to "develop safe, trustworthy, controllable, reliable and scalable AI technology." Bo Zhang (张钹), "Bo ZHANG — Chinese Perspectives on AI Safety," trans. Concordia AI, Chinese Perspectives on AI, accessed October 12, 2023, https://chineseperspectives.ai/Bo-ZHANG.

researchers on technical alignment research, in a speech widely reported in Western media. Both Stuart Russell and Andrew Yao addressed the need to ensure AI decisions reflect universal interests and the importance of achieving international cooperation to prevent an AGI arms race. Geoffrey Hinton, often referred to as the "Godfather of AI," gave a speech on why superintelligence may occur earlier than he previously thought, and the difficulties of ensuring human control. AI researchers Yang Yaodong and Huang Minlie, presented their technical research to improve the safety and alignment of LLMs (refer to the Technical Safety Developments section for more details). Huang Tiejun, previously mentioned in this section, concluded the forum with a speech as the Dean of BAAI. He discussed the difficulties of humans controlling an entity that has surpassed human intelligence, arguing that we know so little about how to build safe AI that discussion must continue.

The trends observed at the 2023 ZGC Forum and the BAAI Conference point to a perceptible shift in the landscape of Chinese expert discussions around frontier AI risks. Historically characterized by skepticism and caution, the discourse at this year's conferences suggest a growing willingness to engage with the topic in a more serious and nuanced manner. Not only did key figures like Robin Li voice concerns related to frontier AI and Kai-Fu Lee discuss frontier AI safety-relevant technical measures, but high-profile events dedicated significant time to addressing extreme AI risks and potential solutions. The fact that frontier risks have found a crucial place at these conferences suggests a broader shift in perception, opening the door for further investigation and dialogue on the subject within China.

Today, many influential Chinese experts voice concerns about extreme risks from AI. Several of the individuals mentioned earlier in this section remain active in this discourse. For example, in July 2023, professor Gao Qiqi published the second iteration of his report on the development and governance of generally capable large models, which features sections on the existential, technical, and governance risks posed by such models.[182] That same month, at the first ever UNSC meeting on AI, Zeng Yi highlighted concerns that AI has "a

[182] "The 'Progress and Response Report of the General Large Models (Version 2.0)' Has Been Officially Released, and Experts Are Discussing 'Knowledge Production and Political Order in the GPT Era' ('通用大模型的进展与应对报告（2.0版）'正式发布 专家热议'GPT时代的知识生产与政治秩序')," Weixin Official Accounts Platform, July 10, 2023, http://mp.weixin.qq.com/s?__biz=MzAxNzAyMzEzNQ==&mid=2649732575&idx=1&sn=bbcf35d4bcd353ed7ad72046964b5626&chksm=83f0fc2fb48775397c0c3d8a9f65e3fd6f8cfc73f1093f85f7e87cb9ae1ec56cf56aea8d5979#rd.

risk of causing human extinction" and called for the creation of a UNSC working group to address AI's challenges.[183]

While this section's central claim is that it has become widely acceptable for experts to discuss frontier AI risks in China without being discredited, this does not mean that the topic is without skeptics. One influential example is Ma Yi, Director & Chair Professor of Musketeers Foundation Institute of Data Science and Department of Computer Science at the University of Hong Kong. In an April 2023 interview, he stated that although "machines [are] beginning to exhibit some human-like reasoning or analytical abilities … I believe this is not based on understanding," and hence "there is no need to worry too much from a technical perspective." However, he allowed that if systems can indeed conduct "autonomous learning," then "people may need to consider some issues in advance."[184] Nonetheless, the mainstream acceptability for Chinese experts to discuss frontier AI risks no doubt paves the way for more nuanced discussions and debates on the topic in future.

Bottom-Line Thinking: A Chinese Perspective on AI Risks

The concept of "bottom-line thinking" was popularized by President Xi Jinping and has been used by Xi and the CPC in a range of contexts, from pandemic preparedness to financial risks.[185] Although the concept lacks a precise definition, it generally emphasizes the identification of worst-case scenarios and red lines and encourages taking preventative measures to avoid their realization.[186] Since its coinage, use of the term has expanded into the wider Chinese intellectual discourse, including with regard to AI risks. Numerous

[183] Concordia AI, "AI Safety in China #1," Substack newsletter, *AI Safety in China* (blog), August 24, 2023, https://aisafetychina.substack.com/p/a6e4bdf0-f687-4ff7-b2c4-dfb06ababd1f.

[184] Yicai (第一财经), "Exclusive Dialogue with Ma Yi, Dean of Mathematics at Hong Kong University: Concern about AI's Domination of the World Is Unfounded and the 'Singularity' Is Far from Coming (独家对话港大数科院长马毅：担心AI统治世界是杞人忧天，'奇点'远未到来)," Sina Finance, April 20, 2023, https://finance.sina.cn/2023-04-20/detail-imyqzcvt5065964.d.html?from=wap.

[185] "Why Does Xi Jinping Emphasize 'Bottom-Line Thinking' (习近平为什么强调'底线思维')," China Daily, January 30, 2019, https://china.chinadaily.com.cn/a/201901/30/WS5c510c53a31010568bdc7697.html.

[186] "Bottom-Line Linking (底线思维)," Sogou Encyclopedia (搜狗百科), accessed October 16, 2023, https://baike.sogou.com/v59638837.htm?fromTitle=%E5%BA%95%E7%BA%BF%E6%80%9D%E7%BB%B4%EF%BC%88%E4%B8%93%E4%B8%9A%E6%9C%AF%E8%AF%AD%EF%BC%89; Guozha Zhang (张国祚), "Discussing 'Bottom-Line Thinking' (谈谈'底线思维')," October 1, 2013, https://news.12371.cn/2013/10/01/ARTI1380592471362492.shtml?from=groupmessage&isappinstalled=0.

Chinese experts—including Jiang Xiaoyuan, He Huaihong, and Li Xiuquan mentioned above—have applied the concept in their works.

For example, in a 2016 article titled "Technological Innovation Should Establish Bottom-Line Thinking—Taking the Development of Artificial Intelligence as an Example," Jiang stated: "In everything, we should prepare for the worst while striving for the best results. Recognizing the immediate, long-term, and ultimate threats of artificial intelligence, we must be cautious when developing AI and not act blindly. Establishing bottom-line thinking is precisely to take action before chaos ensues and to govern before things get out of hand."[187] Similarly, in his book *Does Humanity Still Have A Future?*, He explains that "Bottom-line thinking can be described as a way of thinking that prioritizes considering the worst-case scenario and taking steps to prevent and mitigate the emergence of the worst situation."[188] To apply bottom-line thinking to frontier AI risks, He suggests that developers should refrain from developing general-purpose AI and restrict the "use of violence by robots," among other things.[189] In his 2021 book *The Intelligent Revolution: The Evolution and Value Creation of AI Technology,* Li Xiuquan focuses on maintaining control over "superintelligence" as a key bottom-line, and argues that humanity should not proceed in developing superintelligence until "we have first addressed all control and safety concerns."[190] In a 2022 journal article, **LIU Yidong (刘益东)**, a Chinese scholar with over two decades of experience researching the risks posed by powerful technologies to humanity,[191] wrote about bottom-line thinking in the development of technology more broadly. Liu wrote, "For the future of humanity, the consequences of optimism and pessimism are asymmetrical: The pessimists are worried about technology developing too fast and hope that some fields will pause or slow down. If the pessimists are wrong, at most, the development of technology is slowed or delayed. However, if the

[187] Xiaoyuan Jiang (江晓原), "Scientific and Technological Innovation Should Be Established upon Bottom-Line Thinking - Taking the Development of AI as an Example (科技创新应树立底线思维——以人工智能发展为例)," CPC News, July 29, 2016, http://theory.people.com.cn/n1/2016/0729/c40531-28593493.html.

[188] See chapter on "Artificial Intelligence and Bottom-line Thinking" (人工智能与底线思维): Huaihong He (何怀宏), *Does Humanity Still Have A Future? (人类还有未来吗?)* (Guangxi Normal University Press (广西师范大学出版社), 2020), https://m.douban.com/book/subject/35197706/.

[189] Huaihong He (何怀宏), *Does Humanity Still Have A Future? (人类还有未来吗).*

[190] Xiuquan Li (李修全), *The Intelligent Revolution: The Evolution and Value Creation of AI Technology (智能化变革: 人工智能技术进化与价值创造)* (Tsinghua University Press (清华大学出版社), 2021), https://www.amazon.com/%E6%99%BA%E8%83%BD%E5%8C%96%E5%8F%98%E9%9D%A9-%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E6%8A%80%E6%9C%AF%E8%BF%9B%E5%8C%96%E4%B8%8E%E4%BB%B7%E5%80%BC%E5%88%9B%E9%80%A0-%E6%9D%8E%E4%BF%AE%E5%85%A8/dp/7302578443.

[191] "Liu Yidong (刘益东)," The Institute for the History of Natural Sciences, Chinese Academy of Sciences (中国科学院自然科学史研究所), accessed October 12, 2023, http://sourcedb.ihns.cas.cn/cn/ihnsexport/200906/t20090602_253811.html.

optimists are wrong, it could lead to irreversible consequences, missing the last opportunity for humanity to save itself. Therefore, using bottom-line thinking, the concerns of the pessimists should be taken seriously."[192]

Bottom-line thinking has also been referenced in major domestic governance documents related to AI like the State Council's "Opinions On Strengthening Science and Technology Ethics Governance,"[193] and in international AI governance documents like "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)."[194] The widespread adoption of the concept of bottom-line thinking in AI discourse signals a desire to prioritize safety over rapid, unchecked AI development. Moreover, bottom-line thinking bears similarities with the precautionary principle present in some European Union policies, including the EU AI Act.[195] These similarities may provide common ground for cooperation.

Implications

In conclusion, this section posits that the discourse among Chinese experts on extreme AI risk has shifted from the periphery to the mainstream. Today, a wide range of Chinese scholars across different disciplines have expressed concerns about and are attempting to mitigate these issues. Crucially, the growing engagement of China's leading AI experts offers

[192] Yidong Liu (刘益东), "Comparative Analysis and Evaluation of Two Types of Science and Technology Ethics (对两种科技伦理的对比分析与研判)," People's Tribune (人民论坛), June 2, 2022, https://web.archive.org/web/20220705220542/http://www.rmlt.com.cn/2022/0602/648400.shtml.

[193] "General Office of the Communist Party of China and General Office of the State Council on Publishing 'Opinions On Strengthening Science and Technology Ethics Governance' (中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》)," March 20, 2022, https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

[194] "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)," Ministry of Foreign Affairs (外交部), November 17, 2022, https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html.

[195] Didier Bourguignon, "The Precautionary Principle: Definitions, Applications and Governance," EPRS | European Parliamentary Research Service, September 12, 2015, https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf; Sabine Neschke, Jeremy Pesner, and John Soroushian, "Artificial Intelligence Policy and the European Union: A Look Across the Atlantic" (Bipartisan Policy Center, August 2022), https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2022/08/AI-Policy-and-EU-Final-Paper.pdf; OECD, *Understanding and Applying the Precautionary Principle in the Energy Transition* (Paris: Organisation for Economic Co-operation and Development, 2023), https://www.oecd-ilibrary.org/governance/understanding-and-applying-the-precautionary-principle-in-the-energy-transition_5b14362c-en, pages 86-87.

valuable opportunities for establishing greater international consensus on frontier AI risks and collaborative efforts to address them.

Lab Self-Governance

Takeaways

- Certain Chinese AI labs publish details about alignment procedures, which primarily involve RLHF since 2022, for LLMs they have released. Occasionally, they also release safety evaluations for models focusing on bias and truthfulness.
- Chinese labs have not begun evaluating models for potentially dangerous capabilities before release.
- In September 2023, a leading Chinese AI industry association announced two projects on AI safety/security and AI alignment, demonstrating attention to frontier AI risks.
- From approximately 2018 to 2022, self-governance of AI labs in China primarily consisted of individual labs' promises to adhere to certain ethical principles and those labs' creation of internal ethics review committees.

Safety Practices in Large Model Development

Some Chinese labs publish information on their safety and alignment practices in technical papers released alongside their models. Research on Chinese large-scale pre-trained models from 2020 to 2022 found that, of 26 sampled papers, 12 discussed ethics or governance issues, emphasizing concerns related to bias and fairness, misuse, environmental harms, and utilizing access restrictions.[196]

After the November 2022 launch of GPT-3.5, Chinese lab technical papers have largely continued this trend, focusing on truthfulness and bias when evaluating metrics on safety, and primarily aligning models only using RLHF. These safety measures and metrics could be motivated by multiple factors, and do not necessarily indicate a focus on safety, since they also facilitate models' capabilities. In June, SHLAB and SenseTime released a technical paper accompanying their release of InternLM.[197] According to the paper, InternLM's alignment mirrors that of InstructGPT: it starts with SFT, then reward model training, and finally, RLHF.

[196] Jeffrey Ding and Jenny Xiao, "Recent Trends in China's Large Language Model Landscape" (Centre for the Governance of AI, April 28, 2023), https://cdn.governance.ai/Trends_in_Chinas_LLMs.pdf.
[197] InternLM Team, "InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities," GitHub, June 3, 2023, https://github.com/InternLM/InternLM-techreport/blob/main/InternLM.pdf.

The safety evaluations used by the authors include truthfulness (using TruthfulQA) and bias (using CrowS-Pairs), both of which are widely used benchmarks.

Similarly, Baichuan Intelligence published a technical report alongside the second version of its open-source LLM family, Baichuan 2, in early September.[198] The authors state that Baichuan 2 was aligned through SFT and RLHF. To enhance model safety, the researchers stated that they employed red-teaming, wherein an expert annotation team of ten led a larger "outsourced" team in generating 200,000 attack prompts for safety reinforcement training. They then evaluated safety of the model using the Toxigen dataset and created their own Baichuan Harmless Evaluation Dataset (BHED).

Alibaba's technical report for its open-source Qwen model family, released in August and comparable to InternLM and Baichuan 2, also discussed its usage of SFT and RLHF for alignment.[199] The authors also note that they annotated human-style conversations to improve model helpfulness and claim to have prioritized safety by annotating data related to violence, bias, and pornography. They assert that they evaluated model safety by creating a dataset of 300 instructions for human evaluation and used the MMLU, C-Eval, GSM8K, HumanEval, and BBH benchmarks, though those benchmarks do not appear to be primarily safety-oriented.

Zhipu AI, upon releasing its ChatGLM-6B model in March 2023, included a disclaimer about the model's limitations. The disclaimer noted the model's small capacity, which increased the likelihood of generating incorrect information. It also mentioned that the model had undergone only preliminary alignment with human intentions, and therefore the possibility of the model generating harmful or biased content could not be ruled out. Zhipu AI provided examples of ChatGLM-6B making incorrect statements or being misled.[200] Chinese-LLaMA-Alpaca-2 developers also note similar limitations around potential for their

[198] Baichuan Inc., "Baichuan 2: Open Large-Scale Language Models," Baichuan AI, accessed October 19 2023, https://cdn.baichuan-ai.com/paper/Baichuan2-technical-report.pdf.
[199] Qwen Team, Alibaba Group, "Qwen Technical Report," accessed October 12, 2023, https://qianwen-res.oss-cn-beijing.aliyuncs.com/QWEN_TECHNICAL_REPORT.pdf.
[200] Knowledge Engineering Group (KEG) & Data Mining at Tsinghua University, "ChatGLM-6B/README.Md at Main · THUDM/ChatGLM-6B," GitHub, April 25, 2023, https://github.com/THUDM/ChatGLM-6B/blob/main/README.md.

models to produce "unpredictable harmful content and content that does not conform to human preferences and values," though it does not provide specific examples.[201]

InternLM, Baichuan 2, and ChatGLM-6B do not appear to have been tested for more dangerous capabilities nor have they been subjected to additional alignment procedures beyond RLHF, though Baichuan 2 team claimed to conduct red-teaming. This approach is understandable, given that the model capabilities appear to be on par with GPT-3.5, as suggested by the common Chinese capabilities benchmark SuperCLUE.[202] Therefore, these models may not necessarily require additional alignment procedures or evaluations to achieve a baseline of safety. That said, as Chinese AI capabilities continue to advance, there are opportunities to improve safety through adopting or developing more sophisticated alignment techniques and by testing for dangerous capabilities, such as deceptiveness, power-seeking behavior, and self-duplication. In addition, greater transparency by Chinese AI companies, perhaps through sharing model cards or system cards, would foster trust and awareness in the industry about best practices for safety of frontier models.

Industry Association Actions

Industry-wide action on AI safety has included concrete projects relating to AI safety as well as broad, voluntary principles. The analysis below focuses on China's Artificial Intelligence Industry Alliance (AIIA) (中国人工智能产业发展联盟). This focus is due to the heft of AIIA's membership and the relatively greater transparency regarding its activities. AIIA was founded in 2017 under the direction of four government departments (NDRC, MOST, MIIT, and CAC), led by institutions including MIIT's CAICT research institution.[203] Pan Yunhe, a leading Chinese computer scientist and Academician of the Chinese Academy of Engineering, serves as its chairman. Its Vice-Chairs include a number of representatives from leading universities, research institutions, private companies, and state-owned enterprises.[204] It also possesses an expert committee providing advice on AI governance and Trustworthy

[201] Yiming Cui, "Chinese-LLaMA-Alpaca/README_EN.Md at Main · Ymcui/Chinese-LLaMA-Alpaca," GitHub, April 17, 2023, https://github.com/ymcui/Chinese-LLaMA-Alpaca.
[202] Jeffrey Ding, "ChinAI #231: Latest SuperCLUE Rankings of Large Language Models," Substack newsletter, *ChinAI Newsletter* (blog), July 31, 2023, https://chinai.substack.com/p/chinai-231-latest-superclue-rankings.
[203] "China Artificial Industry Industry Alliance (AIIA) Founded, Institute of Autonomation Selected as a Vice-Chairman Institution (中国人工智能产业发展联盟成立，自动化所当选副理事长单位)," Institute of Automation, Chinese Academy of Sciences (中国科学院自动化研究院), October 13, 2017, http://www.ia.cas.cn/xwzx/ttxw/201710/t20171013_4873206.html.
[204] "About AIIA (关于联盟)," Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), accessed October 15, 2023, http://www.aiiaorg.cn/index.php?m=alliance&c=index&a=structure&s=2#zjjg.

AI.[205] However, AIIA is certainly not the only relevant AI industry association—the China Association for Artificial Intelligence (中国人工智能协会) is another, more academically-oriented association.[206]

As documented in this report's Domestic AI Governance section, as of 2019, AIIA and CAICT had already developed a number of AI evaluations, including for "Trustworthy AI" (可信AI评估) and "content safety" (内容安全评测). AIIA and CAICT also collaborated on a testing system for large models starting in 2021, aiming at the development of standards and tests both for capabilities and also for safety/security and trustworthiness.[207] Then, in September and October 2023, AIIA announced new projects relating to AI safety. On September 27, AIIA announced that CAICT has entrusted it to create a "safety/security governance committee."[208] This body is tasked with channeling industry inputs into policy processes, supporting the development of AI safety/security platforms, promoting learning from the safety supervision industry, and spearheading discussions on AI safety/security governance.[209] Subsequently, on October 8, AIIA announced another project with CAICT called "Deep Alignment,"[210] noting that aligning AI to human values has become "increasingly

[205] "Groups (工作组&推进组&委员会)," Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), accessed October 16, 2023, http://www.aiiaorg.cn/index.php?m=alliance&c=index&a=workgroups&mgroup=3.

[206] "Introduction to the Chinese Association for Artificial Intelligence (中国人工智能学会简介)," Chinese Association for Artificial Intelligence (中国人工智能学会), accessed October 15, 2023, https://www.caai.cn/index.php?s=/home/article/index/id/2.html.

[207] Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Trustworthy AI Technology Hotspot | Large Models Continually Release Technological Dividends, and an Evaluation System for the Extremely Large Model Industry Is Formally Released (可信AI技术热点｜大模型持续释放技术红利，产业级大模型评估体系正式发布)," Weixin Official Accounts Platform, June 27, 2022, http://mp.weixin.qq.com/s?__biz=MzU0MTEwNjg1OA==&mid=2247499125&idx=2&sn=fc677dcdd56cc78b59563798bfedc2c7&chksm=fb2c4ab0cc5bc3a687deebc43d07a53b3e0ac79829fc77a4b8cbd4630824c622c2191005dbf2#rd.

[208] Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Notice on the Preparation for the Establishment of the AIIA Safety/Security Governance Committee: Solicitation for the First Batch of Member Units Simultaneously Starts (关于筹备成立AIIA安全治理委员会的通知首批成员单位同步开始征集)," Weixin Official Accounts Platform, September 27, 2023, http://mp.weixin.qq.com/s?__biz=MzU0MTEwNjg1OA==&mid=2247501263&idx=1&sn=b7dfeb79135ef1f27faaee7137b320f5&chksm=fb2c720acc5bfb1cec2752c16933bbaf92837a432bec282243f65894160 8f1b68f125bcfc7ed#rd.

[209] Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Notice on the Preparation for the Establishment of the AIIA Safety/Security Governance Committee: Solicitation for the First Batch of Member Units Simultaneously Starts (关于筹备成立AIIA安全治理委员会的通知首批成员单位同步开始征集)."

[210] "Deep alignment" is an English name provided by AIIA; the direct translation of the Chinese name is "AI Value Alignment Partnership Plan" (人工智能价值对齐伙伴计划). Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Notice on Preparing to Establish the AIIA 'Artificial Intelligence Value Alignment Partnership Plan' and Soliciting the First Batch of Member Units (关于筹备成立AIIA'人工智能价值对齐伙伴

urgent." As part of this venture, AIIA announced its intent to recruit its first batch of partners and disclosed plans to release a research report titled "AI Values Alignment Operationalization Guide" (人工智能价值对齐操作指南). This endeavor also seeks to promote development of technical tools to evaluate model alignment.[211] Collectively, these projects show AIIA's growing interest in AI safety and security issues. It is unclear how much AIIA will influence the overall AI industry through these actions, but it demonstrates that at least one major AI industry association is paying attention to frontier AI risks.

AIIA and another local AI industry association have, in the past, taken steps to establish joint pledges with ethical principles among AI companies, though this has not been an area of focus in recent years. On May 31, 2019, AIIA released a draft Joint Pledge on Artificial Intelligence Industry Self-Discipline.[212] This document outlined several principles that China's AI companies should follow, including being human-oriented, ensuring safety and controllability, and emphasizing transparency, explainability, and privacy. Significantly, the document also underscored the importance of AI companies helping to formulate standards on AI and educating the public about AI. However, despite plans to publish the final version with the initial list of signatory companies in August 2019, it seems that neither a conclusive version of the final joint agreement nor the signatory companies has been made public.[213] In August 2019, the Shenzhen AI Industry Association also published its own "New Generation AI Industry Self-Discipline Joint Pledge,"[214] which bore resemblances to the AIIA document,

计划'并征集首批成员单位的通知)," Weixin Official Accounts Platform, October 8, 2023, https://mp.weixin.qq.com/s/rzw-zTB2bO34Aeun6oHZ2g.

[211] Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), "Notice on Preparing to Establish the AIIA 'Artificial Intelligence Value Alignment Partnership Plan' and Soliciting the First Batch of Member Units (关于筹备成立AIIA'人工智能价值对齐伙伴计划'并征集首批成员单位的通知)."

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

[213] "The Artificial Intelligence Industry Alliance Signed the 'AI Industry Self-Governance Convention' Proposal (人工智能产业发展联盟签署《人工智能行业自律公约》的倡议)," Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), August 8, 2019, http://www.aiiaorg.cn/index.php?m=content&c=index&a=show&catid=3&id=49; "AIIA Overall Group 2020 Report on the Rationale behind Our Work (AIIA总体组2020年工作思路汇报)," Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟), May 7, 2020, http://www.aiiaorg.cn/index.php?m=content&c=index&a=show&catid=2&id=263.

[214] "Guangdong enterprises take the lead in initiating an AI industry self-discipline convention, covering 8 aspects including privacy protection (粤企领衔发起AI行业自律公约，涉及隐私保护等8个方面)," Nanfang Daily (南方日报), August 19, 2019, https://xapp.southcn.com/node_fb07388412?url=https://law.southcn.com/node_d384a70bd7/7e7932524c.shtml&is_app=1; "Testin Cloud Testing Initiates and Participates in the 'New Generation AI Industry Self-Governance Convention' Assisting the Sustainable Development of the AI Industry (Testin云测发起并参与

emphasizing concepts like human-centered design, fairness, safety, and controllability. Companies including Megvii, iFlytek, Coocaa, and Orbbec were among the signatories.

Lab and Industry Principles

Numerous Chinese labs have issued non-binding statements on AI ethics and governance. These statements range in complexity: some, like Alibaba, offer simple lists of key terms, while others, such as Tencent (illustrated in Appendix E), provide more lengthy explanations of key principles. Some companies have also publicly announced the establishment of AI or S&T ethics committees, which include outside experts, to advise the company on key ethical decisions or codes. We have summarized data on AI ethics principles from a set of key Chinese AI labs in Appendix E. The labs chosen for inclusion in this set were selected based on their prominence in cutting-edge AI research and are not intended to be exhaustive.

Verifying the extent to which companies adhere to their stated principles is challenging. While only a subset of companies have publicly announced the creation of AI or S&T ethics committees—an essential step in institutionalizing ethics reviews—the mere creation of such a committee does not guarantee greater implementation of ethical practices.

Publicly available information is insufficient to allow external commentators to understand the degree to which Chinese labs and researchers have enacted sufficient self-governance measures. Greater transparency would be beneficial. Overall, the principles of these companies largely mirror principles stated in Appendix A, referencing issues such as fairness, protecting personal information, controllability, and benefiting humanity. However, this does not indicate much focus on safety concerns of frontier models apart from the reference to controllability, which, on its own, is a weak indicator.

Compared to voluntary commitments and industry self-governance in other jurisdictions, Chinese companies have stated similar aspirational principles but have not yet pursued concrete commitments. For instance, several leading US companies have thus far promised to conduct internal and external red teaming on risks, create bounty programs for AI vulnerabilities, and deploy AI watermarking technology.[215] Additionally, the Frontier Model

《新一代人工智能行业自律公约》，助力AI行业可持续发展),” China Daily, August 20, 2019, https://tech.chinadaily.com.cn/a/201908/20/WS5d5b8732a31099ab995da7d8.html.

[215] “Voluntary AI Commitments,” White House, September 2023, https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

Forum, which includes leading Western companies Anthropic, Google, Microsoft, and OpenAI, intends to coordinate research on topics such as adversarial robustness, mechanistic interpretability, and emergent behaviors.[216] The forum also aims to share a public library of technical evaluations and benchmarks for frontier models. Chinese labs may not have adopted similar commitments yet because the capabilities of the frontier models are closer to GPT-3.5 levels than GPT-4, and since the private industry in China may be more accustomed to waiting for the government to initiate policy actions. Projects like AIIA's "Deep Alignment" (or subsequent efforts, if they occur) could serve as instruments to facilitate Chinese labs jointly signing on to similar voluntary commitments.

Implications

China's AI industry has room to improve corporate governance, especially once companies develop models that are closer to or even exceed GPT-4 in performance. Companies should continue to release details about the technical alignment of models and performance on safety evaluations. They should additionally consider increasing the variety of alignment techniques used and evaluating models for more dangerous capabilities. Concurrently, companies should consider methods for cooperation on measures such as evaluations or safety research that can reduce AI risks. AIIA or other industry associations also have a potential role to play in this process, especially with recent efforts by AIIA to further research AI safety/security and alignment.

"Ensuring Safe, Secure, and Trustworthy AI," White House, July 2023, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

[216] OpenAI, "Frontier Model Forum," OpenAI, July 26, 2023, https://openai.com/blog/frontier-model-forum.

Public Opinion on AI

Takeaways

- Our takeaways are low-confidence due to the limited amount of good survey research in existence on Chinese public views regarding AI.
- Chinese people appear to generally view the benefits of AI as outweighing the risks.
- Chinese experts and citizens familiar with AI appear to think there is a risk of advanced or Strong AI causing human extinction. However, both groups seem to believe that this risk is controllable.
- Chinese experts and citizens familiar with large models were roughly split on their support for a temporary pause on AI development, as advocated by an FLI open letter.
- The specific policy measures favored by the Chinese population to mitigate AI-associated risks remain ambiguous.

Introduction

The current state of public opinion polling on AI in China, while more extensive than some might expect, remains in an early stage of development. We have compiled a non-exhaustive database of AI-related polling of China (Appendix F), including both Chinese and non-Chinese sources. We then analyze the polling data in terms of three main questions:

1. How do Chinese citizens see the benefits of AI compared to the risks? What are the main risks Chinese people associate with AI?
2. What are Chinese peoples' views on frontier AI risks?
3. What government policies on AI do Chinese people support?

While some observers may downplay the significance of Chinese public opinion for influencing Chinese technology policy, there is precedent for public opinion influencing debate over key legislation and regulation.[217] For example, a lawsuit by a legal professor

[217] Rogier Creemers and Graham Webster, "Translation: Personal Information Protection Law of the People's Republic of China - Effective Nov. 1, 2021," *DigiChina* (blog), August 20, 2021, https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/; China Law Translate, "Translation: Provisions on the Management of Algorithmic

against a Hangzhou wildlife park over usage of facial recognition technology created public controversy and triggered debate over privacy rights, which influenced discussions around the Personal Information Protection Law.[218] Journalism investigating exploitation of Chinese food delivery drivers through algorithms was a major factor influencing Chinese regulation over recommendation algorithms.[219] However, foreign observers should not assume that these debates directly mirror those in other jurisdictions.

Unfortunately, of the six relevant surveys we could find, five pre-dated 2023. As discussed in earlier sections of this report, we believe there has been a significant increase in global and Chinese policy awareness and expert buy-in on frontier AI risks over the course of late 2022 and 2023. It is highly plausible that this shift in awareness and buy-in has also affected public opinion, but our survey sample would mostly fail to capture such an update. Moreover, existing surveys on AI risks lack precision and rigor. None claim to be representative samples of the Chinese population, and most do not explicitly ask about frontier or extreme risks. They were all online surveys, with some distributed privately via WeChat, which resulted in biased samples–often skewed towards more educated respondents or those disproportionately in the AI field. It seems researchers did not adjust for these biases. In addition, some surveys are conducted by state media sources or involve private tech companies, rather than professional polling companies, further calling results into question. Therefore, one should interpret these findings with caution. Some may question the reliability of any poll in China, no matter how representative, due to extensive government involvement in citizens' lives which may lead respondents to self-censor. However, there is substantial precedent in the academic and research community for analyzing Chinese public opinion surveys, providing they are well-designed.[220] Therefore, we

Recommendations in Internet Information Services," *China Law Translate* (blog), January 4, 2022, https://www.chinalawtranslate.com/algorithms/.

[218] "Reframing AI Governance: Perspectives from Asia," Digital Futures Lab | Konrad-Adenauer-Stiftung, July 2022, 78, https://assets.website-files.com/62c21546bfcfcd456b59ec8a/62df3bbcd1d3f82534a706f1_%E2%80%A2Report_AI_in_Asia.pdf.

[219] Matt Sheehan and Sharon Du, "How Food Delivery Workers Shaped Chinese Algorithm Regulations," Carnegie Endowment for International Peace, November 2, 2022, https://carnegieendowment.org/2022/11/02/how-food-delivery-workers-shaped-chinese-algorithm-regulations-pub-88310.

[220] Dan Harsha, "Long-Term Survey Reveals Chinese Government Satisfaction," *The Harvard Gazette* (blog), July 9, 2020, https://news.harvard.edu/gazette/story/2020/07/long-term-survey-reveals-chinese-government-satisfaction/; Reza Hasmath, "How China Sees the World in 2023" (The China Institute, University of Alberta, May 2023), https://www.ualberta.ca/china-institute/media-library/media-gallery/research/research-papers/2023-china-survey-report/howchinaseestheworld2023.pdf; Ilaria Mazzocco and Scott Kennedy, "Public Opinion in China: A

believe that surveys on AI risks by independent and experienced polling organizations would be beneficial for better understanding how the Chinese public thinks about AI risks.

Weighing AI Benefits Against AI Risks

Overall, Chinese respondents expressed positive views about AI development, while registering concerns about several specific risks. For instance, the 2021 Ipsos survey found that 78% of respondents in China thought that AI products and services "have more benefits than drawbacks." This was the highest among all 28 countries surveyed and notably exceeded the overall average of 52%. In the same survey, Chinese respondents also had the highest level of trust in companies that use AI and a low level of nervousness about using AI products (4th lowest among 28 countries).[221] A 2019 poll by Chinese tech company Cheetah Mobile, reflecting a younger and more rural population, found low negative feelings about AI development (worried, 担忧, 9%, skeptical, 质疑, 5%, anxious, 焦虑, 3%) compared to positive feelings (expectant, 期待: 61%, excited, 兴奋: 56%).[222] The 2023 CLAI survey also found that 86% of respondents familiar with large AI models thought that the impact of "continuous research and application of large AI models on society" is positive.[223]

While these polls suggest that the Chinese population is overall more enthusiastic about the benefits of AI than concerned about its risks, one should interpret these findings cautiously given their methodological limitations, including participation of tech companies incentivized to skew the results in a positive direction. Excluding the Ipsos poll, which is just one study, it is difficult to directly draw comparisons with polling conducted in other countries due to concerns around consistency and methodology.

Liberal Silent Majority?," February 9, 2022, https://www.csis.org/analysis/public-opinion-china-liberal-silent-majority.

[221] "Global Opinions and Expectations About Artificial Intelligence: A Global Advisor Survey" (Ipsos, January 2022), https://www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Global-opinions-and-expectations-about-AI-2022.pdf.

[222] Cheetah User Research Center (猎豹用户研究中心), "Cheetah Announcement | AI in the Eyes of Ordinary People: A Research Report on Public Recognition, Feelings, and Attitudes towards AI (豹告 | 普通人眼中的AI：大众AI认知、感受、态度调研报告)," Weixin Official Accounts Platform, September 9, 2019, http://mp.weixin.qq.com/s?__biz=MzU3NjI0MzgxOQ==&mid=2247489828&idx=1&sn=2efbbed9c94c8fbb98e5fcbc2700417c&chksm=fd178820ca6001361dc29bacfd925439e06505b3efdcd8475d8a738c21028ac0a5a628a21d2f#rd.

[223] Yi Zeng (曾毅) et al., "Voices from China on 'Pause Giant AI Experiments: An Open Letter,'" Center for Long-term Artificial Intelligence, April 4, 2023, https://long-term-ai.center/research/f/voices-from-china-on-pause-giant-ai-experiments-an-open-letter.

Privacy and data security, economic displacement, and overreliance on AI are some of the key AI risks that concern the Chinese public. The 2021 Ipsos survey found that Chinese people thought the areas of life least likely to improve due to AI were economic factors: 52% felt employment would not improve, and 46% thought the same for the cost of living.[224] A Tsinghua University Center for International Security and Strategy (CISS) report in 2019 found that Chinese youth were most worried about risks of AI in employment (52%) and privacy and ethics (43%).[225] A 2017 Communist Youth Daily poll found that the main risks people identified with AI were overreliance on AI (63%) and privacy (53%).[226] However, none of these polls appear to have included frontier AI risks or existential risks from AI as options.

Frontier AI Risks

To our knowledge, only two surveys have been conducted in China on how Chinese people view frontier AI risks and how to reduce those risks. Both surveys were conducted by the Center for Long-term Artificial Intelligence, which is headed by leading Chinese AI ethics academic Zeng Yi. A March 2023 publication, based on 2021 polling, addressed the development of Strong AI, which the authors defined as "The combination of Artificial General Intelligence/Human-Level AI and Superintelligence." The second survey, released in April 2023, was based on polling that same month, and focused on FLI's "Pause Giant AI Experiments: An Open Letter."[227]

[224] "Global Opinions and Expectations About Artificial Intelligence: A Global Advisor Survey" (Ipsos, January 2022), https://www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Global-opinions-and-expectations-about-AI-2022.pdf, page 13.

[225] "Pre-Research Report: Risks and Governance of AI from the Perspective of Chinese Youth" (Center for International Security and Strategy Tsinghua University (清华大学战略与安全研究中心), 2019), https://ciss.tsinghua.edu.cn/upload_files/atta/1589025522890_83.pdf, pages 8-9.

[226] "77.8% of Respondents Are Optimistic about the Development of Artificial Intelligence in China (77.8%受访者看好我国人工智能的发展)," Communist Youth Daily (中国青年报), June 13, 2017, https://zqb.cyol.com/html/2017-06/13/nw.D110000zgqnb_20170613_1-07.htm.

[227] Yi Zeng (曾毅) and Kang Sun, "Whether We Can and Should Develop Strong AI: A Survey in China," Center for Long-term Artificial Intelligence, Center for Long-term Artificial Intelligence, March 12, 2023, https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence; Yi Zeng (曾毅) et al., "Voices from China on 'Pause Giant AI Experiments: An Open Letter,'" Center for Long-term Artificial Intelligence, April 4, 2023, https://long-term-ai.center/research/f/voices-from-china-on-pause-giant-ai-experiments-an-open-letter.

The survey on strong AI had two main samples: 1) a group of 63 computer science experts[228] who were invited to participate, and 2) 1,032 respondents (including the aforementioned 63 experts), mainly involving "young and middle-aged AI-related students and scholars, as well as participants from other fields." Given the greater relevance of the 63 computer science experts to questions on AI capabilities and risks, as well as a less refined sample for the larger group, we have chosen to focus on the results from the 63 experts. However, results for the larger pool of respondents are quite similar. The 63 experts believed that "Strong AI poses existential risks" to humans (72%), while still arguing that "Strong AI should be developed" (79%), and believing that "Strong AI can coexist harmoniously with human *[sic]*" (69%). At the same time, most of the experts (81%) believe that "achieving Strong AI is possible" but see it as unlikely to occur by 2030 (only 5% think AI will reach human levels by 2030). This survey demonstrates that the Chinese AI research community may have concerns about AI existential risks. However, the interest in developing Strong AI despite awareness of existential risks suggests that respondents generally think the risks are controllable. This is an initial, low-confidence hypothesis that will need to be further refined with future data.

Regarding the survey on the proposition to pause giant AI experiments, roughly an equal number of respondents supported a pause as did not support a pause (33% and 31% respectively), while another 30% thought a pause would be "useless." Reasons cited in favor of a pause encompassed concerns about misuse, threats to human existence, and employment. Drawing definitive conclusions from this survey about how respondents weigh benefits and risks of frontier AI development is challenging, especially given uncertainty among about a third of respondents towards effectiveness of a pause. However, it does suggest that a minority (31%) of informed Chinese (92% of whom were familiar with the term "giant AI models") have substantial concerns about negative societal impacts of frontier AI development. By comparison, Rethink Priorities estimates that, based on a YouGov poll, 51% of the US public would support a worldwide pause on the development of large-scale AI systems for at least six months, while 25% would oppose such a pause.[229]

[228] Fellows or distinguished scholars from the Chinese Association for Artificial Intelligence (CAAI), China Computer Federation (CCF), Chinese Association of Automation (CAA), and Beijing Academy of Artificial Intelligence (BAAI).

[229] Jamie Elsey and David Moss, "US Public Opinion of AI Policy and Risk," Rethink Priorities, May 12, 2023, https://rethinkpriorities.org/publications/us-public-opinion-of-ai-policy-and-risk.

Policy Interventions

Few surveys address potential policy interventions to improve AI's societal impact. In the 2019 CISS poll, youth were asked "How should AI-related risks be prevented and regulated?" and only two answers received over 50% support: passing (unspecified) laws, and promoting vocational education to stimulate employment transformation. There was also limited support for increasing public awareness (just over 35%), and minimal support for reducing R&D, maintaining the status quo, or taxing AI companies.[230] Meanwhile, approximately 91% of respondents in the 2023 survey overseen by Dr. Zeng Yi said that they support implementation of "ethics, safety and governance framework for every large AI model used in social services." Overall, more work is necessary to understand what policies the Chinese public favors on reducing AI risks.

Implications

Surveys of the Chinese public and scientific communities regarding AI indicate that Chinese respondents generally view the benefits of AI quite positively. While these groups have some awareness of existential AI risks, they tend to be optimistic about the possibility of reducing those risks. Limited evidence exists on their policy preferences related to AI. However, these conclusions are preliminary due to the lack of data available, and more work is necessary to understand the views of the Chinese public and of AI researchers on frontier AI risks—Professor Zeng Yi's work is the only work in this space. However, incidents like the public outcry over Chinese food delivery drivers' exploitation through algorithms suggest public opinion could still influence the development of AI policy and safety in China.[231]

[230] "Pre-Research Report: Risks and Governance of AI from the Perspective of Chinese Youth" (Center for International Security and Strategy Tsinghua University (清华大学战略与安全研究中心), 2019), https://ciss.tsinghua.edu.cn/upload_files/atta/1589025522890_83.pdf, pages 12-13.

[231] Matt Sheehan and Sharon Du, "How Food Delivery Workers Shaped Chinese Algorithm Regulations," Carnegie Endowment for International Peace, November 2, 2022, https://carnegieendowment.org/2022/11/02/how-food-delivery-workers-shaped-chinese-algorithm-regulations-pub-88310.

Conclusion

The Chinese government, academics, and technical community are more actively trying to reduce risks from frontier AI models than many foreign observers might realize. It is our understanding that China possesses robust tools for domestic governance, including algorithm registries, safety/security reviews, ethics reviews, and standards. Although these tools have not been extensively utilized to address frontier AI risks, they have the potential to enhance AI safety. Recent documents also suggest that government institutions are increasingly interested in the risks of advanced AI and aligning AI with human values. On the international front, we detailed China's increasingly proactive stances. Our findings indicate that China is concerned about loss of control over military AI systems and general AI systems. China prioritizes engagement through the UN and champions the interests of the Global South. We also analyzed technical safety research in China, spotlighting work on alignment, robustness, and evaluations in various Chinese universities and tech companies. A growing number of influential Chinese experts have voiced concerns about frontier AI risks, especially since the introduction of GPT-3.5. Moreover, Chinese labs have included safety measures and evaluations in their LLMs released in 2023. However, these measures appear less suited to models that exceed GPT-4 in capabilities, and the industry has not yet made new voluntary commitments to ensure safety of large models. We also reviewed the limited literature on Chinese public opinion surveys concerning AI, a field that indeed requires more extensive exploration.

Frontier AI developments have the potential to pose serious and even catastrophic threats to the entire globe. We hope that this report has given outside observers a better appreciation for the level of thinking and action on AI safety occurring in Chinese academia, industry, and government. We exhort all who have an opportunity to cooperate and help address these concerns, and we believe that various Chinese actors have an important role to play in that effort. AI is not the only potentially existential threat humanity has faced; nuclear weapons are likely the most prominent existential threat in the mind of most readers. Early in the Cold War, Albert Einstein and Bertrand Russell wrote an open letter on the dangers of nuclear war, co-signed by nine other prominent intellectuals. We close with their words of wisdom, in hopes that humanity learns the best lessons and avoids the worst mistakes from the world's ongoing attempts to avoid nuclear catastrophe:

> “There lies before us, if we choose, continual progress in happiness, knowledge, and wisdom. Shall we, instead, choose death, because we cannot forget our quarrels? We appeal as human beings to human beings: Remember your humanity, and forget the rest. If you can do so, the way lies open to a new Paradise; if you cannot, there lies before you the risk of universal death.”[232]

[232] “Russell-Einstein Manifesto,” *The National Museum of Nuclear Science & History* (blog), accessed October 13, 2023, https://ahf.nuclearmuseum.org/ahf/key-documents/russell-einstein-manifesto/.

Acknowledgments

We would like to thank (in alphabetical order) Qiqi GAO (高奇琦), David Krueger, Hui LI (李辉), Nicolas Mialhe, Matt Sheehan, and Yi ZENG (曾毅) for their in-depth feedback on the report. All remaining errors are our own.

Appendices

Appendix A. Mentioned Domestic Governance Documents

(In chronological order of final issue date)

Lead author	Date	Document
Li Keqiang, Premier of the State Council	March 5, 2017	Report on the Work of the Government[233]
State Council	July 20, 2017	New Generation AI Development Plan[234]
Industrial Department II of the Standardization Administration of China China Electronics Standardization Institute	January 2018	Artificial Intelligence Standardization White Paper (2018 Edition)[235]
Cyberspace Administration of China	November 15, 2018	Regulations for the Security Assessment of Internet Information Services Having Public Opinion Properties or Social Mobilization Capacity[236]
National New Generation AI Governance Expert Committee	June 17, 2019	Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence[237]
National Information Security Standardization Technical	October 2019	White Paper on AI Safety/Security Standardization[238]

[233] Li Keqiang, "Report on the Work of the Government."

[234] State Council, "Notice by the State Council on the Publication of the New Generation AI Development Plan (国务院关于印发新一代人工智能发展规划的通知)."

[235] Industrial Department II of the Standardization Administration of China (国家标准委工业二部) and China Electronics Standardization Institute, "Artificial Intelligence Standardization White Paper (2018 Edition) (人工智能标准化白皮书)."

[236] "Regulations for the Security Assessment of Internet Information Services Having Public Opinion Properties or Social Mobilization Capacity (具有輿论属性或社会动员能力的互联网信息服务安全评估规定)."

[237] National New Generation AI Governance Expert Committee (国家新一代人工智能治理专业委员会), "Develop Responsible AI: New Generation AI Governance Principles Published (发展负责任的人工智能：新一代人工智能治理原则发布)."

[238] National Information Security Standardization Technical Committee's (TC260) Big Data Security Special Working Group (全国信息安全标准化技术委员会大数据安全标准特别工作组), "AI Safety/Security Standardization White Paper - 2019 Version (人工智能安全标准化白皮书 - 2019 版)."

Committee's (TC260) Big Data Security Special Working Group		
Standardization Administration of China Cyberspace Administration of China National Development and Reform Commission Ministry of Science and Technology Ministry of Industry and Information Technology	August 9, 2020	National New Generation AI Standards System Construction Guide[239]
China Academy of Information and Communications Technology JD Explore Academy	July 2021	White Paper on Trustworthy Artificial Intelligence[240]
National Artificial Intelligence Standardization Overall Group National Information Technology Standardization Committee AI Subcommittee (TC28/SC42) China Electronics Standardization Institute	July 19, 2021	White Paper on Artificial Intelligence Standardization (2021)[241]
National New Generation AI Governance Expert Committee	September 26, 2021	Ethical Norms for New Generation Artificial Intelligence[242]
Cyberspace Administration of China	Published: January 4, 2022 Took effect: March 1, 2022	Administrative Provisions on Algorithm Recommendation for Internet Information Services[243]

[239] Standardization Administration of China (国家标准委) et al., "Notice by Five Departments on the Publication of the 'Standardization Construction Guide for National New Generation AI' (五部门关于印发《国家新一代人工智能标准体系建设指南》的通知)."

[240] China Academy of Information and Communications Technology (CAICT) (中国信通院) and JD Explore Academy (京东探索研究院), "Trustworthy AI White Paper."

[241] Center for Security and Emerging Technology, "Translation: Artificial Intelligence Standardization White Paper (2021 Edition)."

[242] Ministry of Science and Technology (科技部), 'Ethical Norms for New Generation AI' Published (《新一代人工智能伦理规范》发布)."

[243] Cyberspace Administration of China (网信办) et al., "Administrative Provisions on Algorithm Recommendation for Internet Information Services (互联网信息服务算法推荐管理规定)." Cyberspace Administration of China (网信办), January 4, 2022, http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

Ministry of Industry and Information Technology Ministry of Public Security State Administration for Market Regulation		
General Office of the Communist Party of China General Office of the State Council	March 20, 2022	Opinions On Strengthening Science and Technology Ethics Governance[244]
Standing Committee of the Shenzhen Municipal People's Congress	Published: September 9, 2022 Took effect: November 1, 2022	Shenzhen Special Economic Zone AI Industry Promotion Measures[245]
Standing Committee of the Shanghai Municipal People's Congress	October 1, 2022	Regulations for Promoting the Development of the Artificial Intelligence Industry in Shanghai[246]
Cyberspace Administration of China	November 18, 2022	Personal Information Protection Certification Implementation Rules[247]
Cyberspace Administration of China Ministry of Industry and Information Technology Ministry of Public Security	Published: November 25, 2022 Took effect: January 10, 2023	Provisions on Management of Deep Synthesis in Internet Information Service[248]
National Artificial Intelligence Standardization Overall Group National Information Technology Standardization Committee Artificial Intelligence Subcommittee (TC28/SC42)	March 2023	AI Ethics Governance Standardization Guide, 2023 Version[249]

[244] “General Office of the Communist Party of China and General Office of the State Council on Publishing ‘Opinions On Strengthening Science and Technology Ethics Governance’ (中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》).”

[245] “Shenzhen Special Economic Zone AI Industry Promotion Measures (深圳经济特区人工智能产业促进条例).”

[246] “Regulations for Promoting the Development of the Artificial Intelligence Industry in Shanghai (上海市促进人工智能产业发展条例).”

[247] “Personal Information Protection Certification Implementation Rules (个人信息保护认证实施规则).”

[248] Cyberspace Administration of China (网信办), Ministry of Industry and Information Technology (工信部), and Ministry of Public Security (公安部), “Provisions on Management of Deep Synthesis in Internet Information Service (互联网信息服务深度合成管理规定).”

[249] National AI Standardization Overall Group (国家人工智能标准化总体组) and National Information Technology Standardization Committee AI Subcommittee (全国信标委人工智能分委会), “AI Ethics Governance Standardization Guide, 2023 Version (人工智能伦理治理标准化指南, 2023 版).”

The Political Bureau of the CPC Central Committee	April 28, 2023	Analyzing and Researching the Present Economic Situation and Economic Work[250]
Beijing Economics and Informatization Bureau Beijing Science and Technology Bureau Zhongguancun Management Committee Beijing Development and Reform Bureau	May 19, 2023	Beijing Artificial General Intelligence Industry Innovation Partnership Plan[251]
National Information Security Standardization Technical Committee’s (TC260) Big Data Security Special Working Group	May 29, 2023	White Paper on Artificial Intelligence Safety/Security Standardization[252]
General Office of Beijing Municipal People's Government Zhongguancun Management Committee	May 30, 2023	Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing[253]
Cyberspace Administration of China National Development and Reform Commission Ministry of Education Ministry of Science and Technology Ministry of Industry and Information Technology Ministry of Public Security	Published: July 13, 2023 Took effect: August 15, 2023	Interim Measures for the Management of Generative Artificial Intelligence Services[254]

[250] Politburo of the Communist Party of China (中共中央政治局), “Analyzing and Researching the Present Economic Situation and Economic Work (分析研究当前经济形势和经济工作).”

[251] Beijing Economics and Informatization Bureau (北京市经济和信息化局), “Beijing Artificial General Intelligence Industry Innovation Partnership Plan (北京市通用人工智能产业创新伙伴计划).”

[252] National Information Security Standardization Technical Committee’s (TC260) Big Data Security Special Working Group (全国信息安全标准化技术委员会大数据安全标准特别工作组), “AI Safety/Security Standardization White Paper - 2023 Version (人工智能安全标准化白皮书 - 2023 版).”

[253] General Office of Beijing Municipal People’s Government (北京市人民政府办公厅), “Notice by the General Office of the Beijing Municipal People’s Government on the Publication of the ‘Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing’ (北京市人民政府办公厅关于印发《北京市促进通用人工智能创新发展的若干措施》的通知).”

[254] Cyberspace Administration of China (国家互联网信息办公室) et al., “Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法),” Cyberspace Administration of China (网信办), July 13, 2023, http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

National Radio and Television Administration		
Chengdu Municipal Bureau of Economic and Information Technology Chengdu New Economic Development Commission	August 4, 2023	Several Measures for Accelerating Large Model Innovation and New Applications to Promote the High Quality Development of the AI Industry in Chengdu Municipality[255]
State Administration for Market Regulation Standardization Administration of China National Information Security Standardization Technical Committee (TC260)	Published: August 6, 2023 Taking effect: March 1, 2024	Information security technology-Assessment specification for Machine learning algorithms[256]
Research Group for "Investigation on the Status of the Construction of China's Artificial Intelligence Ethics Review and Regulatory System"	August 15, 2023	Artificial Intelligence Law Model Law Version 1.0 (Expert Suggestion Draft)[257]
Ministry of Science and Technology Ministry of Education Ministry of Industry and Information Technology Ministry of Agriculture and Rural Affairs National Health Commission Chinese Academy of Sciences	Published: October 8, 2023 Taking effect: December 1, 2023	Science and Technology Ethics Review Plan (Trial)[258]

[255] Chengdu Municipal Bureau of Economic and Information Technology (成都市经济和信息化局) and Chengdu New Economic Development Commission (成都市新经济发展委员会), "Notice on the Publication of 'Several Measures for Accelerating Large Model Innovation and New Applications to Promote the High Quality Development of the AI Industry in Chengdu Municipality' (关于印发《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》的通知)."

[256] State Administration of Market Regulation (市场监管总局) and Standardization Administration of China (国家标准委), "Information Security Technology-Assessment Specification for Security of Machine Learning Algorithms (GB/T 42888-2023) (信息安全技术-机器学习算法安全评估规范)," August 6, 2023, http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=E7170BA58AE37AACF4170242EFD25183.

[257] New Governance (新治理), "'AI Law (Model Law) 1.0' (Expert Suggestion Draft) Drafting Statement and Full Document (《人工智能法（示范法）1.0》（专家建议稿）起草说明和全文)."

[258] Ministry of Science and Technology (科技部) et al., "Notice on the Publishing of the 'Science and Technology Ethics Review Plan (Trial)' (关于印发《科技伦理审查办法（试行）》的通知)."

Chinese Academy of Social Sciences Chinese Academy of Engineering China Association for Science and Technology Central Military Commission Science and Technology Committee		

Appendix B. Key Documents on International AI Governance Involving China

Author	Date	Document
Chinese delegation to UN Convention on Certain Conventional Weapons (CCW)	December 2016	The position paper submitted by the Chinese delegation to CCW 5th Review Conference[259]
G20	June 2019	G20 AI Principles[260]
UNESCO	November 2021	Recommendation on the Ethics of Artificial Intelligence[261]
Ministry of Foreign Affairs	December 2021	Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)[262]
Ministry of Foreign Affairs	November 2022	Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)[263]
Ministry of Foreign Affairs	February 2023	The Global Security Initiative Concept Paper[264]
Permanent Representative of China to the United Nations Zhang Jun	July 2023	Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security[265]
President Xi Jinping	August 2023	Seeking Development Through Solidarity and Cooperation and Shouldering Our

[259] United Nations Office at Geneva, "The Position Paper Submitted by the Chinese Delegation to CCW 5th Review Conference."

[260] "G20 AI Principles."

[261] UNESCO, "Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library."

[262] Ministry of Foreign Affairs (外交部), "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)."

[263] " United Nations Office at Geneva, "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)."

[264] "The Global Security Initiative Concept Paper."

[265] Jun Zhang (张军), "Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security."

		Responsibility for Peace[266]
Ministry of Foreign Affairs	September 2023	Proposal of the People's Republic of China on the Reform and Development of Global Governance[267]
Cyberspace Administration of China	October 2023	Global AI Governance Initiative[268]

[266] Xi Jinping (习近平), “Full Text: Remarks by Chinese President Xi Jinping at the 15th BRICS Summit.”
[267] Xi Jinping (习近平), “Full Text: Proposal of the People’s Republic of China on the Reform and Development of Global Governance.”
[268] Cyberspace Administration of China (网信办), “Global AI Governance Initiative (全球人工智能治理倡议).”

Appendix C. Key AI Safety Research Labs and Groups in China

Note: These are ordered alphabetically. These labs vary in size substantially, and their research directions may not remain fully consistent. In the future, it is possible that some may switch between different AI safety research directions, and others may enter or leave the AI safety field altogether.

Alibaba DAMO Academy Language Technologies Lab

- Mission: Alibaba DAMO Academy "is dedicated to exploring the unknown through scientific and technological research and innovation," pursuing "the betterment of humanity."[269]
- Key Personnel: **HUANG Fei (黄非)** is the Chief Scientist and Senior Director of the Language Technologies Lab, DAMO Academy, since 2018. He previously worked at Meta (then Facebook) and IBM and received his PhD from Carnegie Mellon University in Language and Information Technologies.[270]
- Key Research Results:
    - RRHF: a more efficient method to align language models with human feedback without complex hyperparameter tuning. April 2023.[271]
    - CVALUES: assesses value alignment in Chinese-language models in terms of both "safety" – level of harmful or risky content in responses – as well as "responsibility" – providing "positive guidance and humanistic care." July 2023.[272]

Fudan University Natural Language Processing Group

- Mission: The Fudan University Natural Language Processing Lab is one of the earliest labs in China to research natural language processing. The lab conducts research in

[269] "About - DAMO Academy," Alibaba DAMO Academy, accessed October 12, 2023, https://damo.alibaba.com/about/.

[270] "Fei Huang," accessed October 12, 2023, https://sites.google.com/view/fei-huang.

[271] Yuan et al., "RRHF."

[272] Guohai Xu et al., "CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility" (arXiv, July 18, 2023), http://arxiv.org/abs/2307.09705.

the fields of neural networks, QA systems, information retrieval, and information extraction.[273]

- Key Personnel: The lab is led by **HUANG Xuanjing (黄萱菁)**, professor in the School of Computer Science at Fudan University. Dr. Huang received a PhD in Computer Science from Fudan (1998).[274] Professor Qiu Xipeng also works at the lab.
- Key Research Results:
  - How Robust is GPT-3.5 to Predecessors: tests robustness of GPT-3.5 using 21 datasets across 9 popular NLP tasks, noting specific robustness challenges including robustness instability, prompt sensitivity, and number sensitivity. March 2023.[275]
  - TextFlint: a multilingual robustness evaluation platform for NLP tasks, including universal text transformation, task-specific transformation, adversarial attack, subpopulation, and combinations therein. May 2021.[276]

Hong Kong University of Science and Technology (HKUST) FU Jie Research Team

- Mission: HKUST visiting scholar **FU Jie (付杰)** is "working towards safe, scalable system-2, mixed-modal interactive LLMs that are capable of observing, acting, and receiving feedback iteratively from external entities."[277] His research on safe and scalable system-2 language models includes work on human-AI alignment, meta learning, and modular neural architectures.
- Key Personnel: Dr. Fu is a visiting scholar at HKUST. He was previously a researcher at BAAI and received his PhD from the National University of Singapore.
- Key Research Results:
  - Interactive Natural Language Processing (iNLP): proposes a novel framework in NLP that can interact with humans, knowledge bases, models, and

[273] "The Fudan Lab For Natural Language Processing Group," The Fudan Lab For Natural Language Processing Group, accessed October 12, 2023, https://nlp.fudan.edu.cn/nlpen/main.htm.
[274] Xuanjing Huang (黄萱菁), "About Me," Xuanjing Huang, accessed October 12, 2023, https://xuanjing-huang.github.io/.
[275] Xuanting Chen et al., "How Robust Is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks" (arXiv, March 1, 2023), https://doi.org/10.48550/arXiv.2303.00293.
[276] Gui et al., "TextFlint."
[277] Jie Fu (付杰), "Big AI Dream | Jie Fu," Big AI Dream | Jie Fu, accessed October 12, 2023, https://bigaidream.github.io/.

  - environment, which the authors argue addresses limitations in alignment. May 2023, authored before Dr. Fu joined HKUST.[278]
  - Chinese Open Instruction Generalist (COIG) Dataset: a Chinese instruction tuning corpus, including 200K Chinese instruction tuning samples and a 3k Chinese human-value alignment corpus. The COIG dataset is available on Hugging Face.[279] April 2023, authored before Dr. Fu joined HKUST.[280]

Huawei Noah's Ark Lab Speech and Language Computing Group

- Mission: Huawei's Noah's Ark Lab is located in Huawei's 2012 Lab and seeks to push the frontier of R&D in AI and big data.[281] The primary research directions are computer vision, decision making & reasoning, speech and language processing, AI theory, search & recommendation, and human-computer interaction.[282]
- Key Personnel: **LIU Qun (刘群)**, Chief Scientist of Speech and Language Computing at Huawei Noah's Ark Lab.[283] Dr. Liu was previously a professor at Dublin City University and at the Chinese Academy of Sciences Institute of Computing Technology. He received a PhD in Computer Science from PKU (2004).
- Key Research Results:
  - Aligning Large Language Models with Human: a survey overviewing techniques for aligning LLMs with human expectations, in terms of data collection, training methodologies, and model evaluation. July 2023.[284]
  - MoralDial (with Tsinghua CoAI, see CoAI section for more details). May 2023.[285]

[278] Zekun Wang et al., "Interactive Natural Language Processing" (arXiv, May 22, 2023), http://arxiv.org/abs/2305.13246.
[279] "BAAI/COIG-PC · Datasets at Hugging Face," October 5, 2023, https://huggingface.co/datasets/BAAI/COIG-PC.
[280] Ge Zhang et al., "Chinese Open Instruction Generalist: A Preliminary Release" (arXiv, April 24, 2023), http://arxiv.org/abs/2304.07987.
[281] Lao Wang (老王), "China's Most Secret Research Base-Huawei 2012 Lab (中国最神秘的研究基地——华为2012实验室)," August 10, 2016, https://www.leiphone.com/category/industrynews/5fWci6bJoL7JW5Wr.html. "About," Huawei Noah's Ark Lab, accessed October 12, 2023, https://noahlab.com.hk/#/about.
[282] "Research," Huawei Noah's Ark Lab, accessed October 12, 2023, https://noahlab.com.hk/#/research.
[283] Qun Liu (刘群), "LIU, Qun (刘群)," accessed October 12, 2023, https://liuquncn.github.io/index_en.html.
[284] Wang et al., "Aligning Large Language Models with Human."
[285] Sun et al., "MoralDial."

Microsoft Research Asia (MSRA) Societal AI Group

- Mission: Microsoft Research Asia has a team working on "Societal AI" (社会责任人工智能), which focuses on ensuring fairness, reliability and safety, privacy and assurance, tolerance, and transparency and responsibility of technical and large model applications.[286]
- Key personnel: The group is led by Xie Xing, a Senior Principal Research Manager at Microsoft Research Asia. He received his PhD in Computer Science from the University of Science and Technology of China (2001).[287] Dr. Xie has noted that one of the five research directions of his team working on "Societal AI" is alignment with human values.[288]
- Key research results:
  - From Instructions to Intrinsic Human Values: surveys alignment goals of previous AI alignment work, categorizing them into three increasingly difficult levels of alignment goals—aligning AI to human instructions, human preferences, and human values. September 2023.[289]
  - PromptBench: a robustness benchmark measuring resilience of LLMs to adversarial prompts, with 4,032 adversarial prompts over 10 tasks and 15 datasets. August 2023.[290]

PKU Alignment and Interaction Research Lab (PAIR Lab)

- Mission: PAIR Lab at Peking University is undertaking research on decision making, strategic interactions, and value alignment. PAIR Lab's alignment research focuses on reinforcement learning from human feedback (RLHF), multi-agent alignment, self-alignment, and constitutional AI, "to steer AGI development towards a safe, beneficial future aligned with the progression of humanity."[291]

[286] "Xie Xing: conduct research that can withstand the test of time, develop responsible AI (谢幸：做经得起时间检验的研究，打造负责任的人工智能)," *MSRA* (blog), August 9, 2023, https://www.msra.cn/zh-cn/news/people-stories/xing-xie-societal-ai.

[287] "Xing Xie at Microsoft Research," *Microsoft Research* (blog), accessed October 12, 2023, https://www.microsoft.com/en-us/research/people/xingx/. Also see "Responsible AI Workshop: An Interdisciplinary Approach," *Microsoft Research* (blog), October 24, 2022, https://www.microsoft.com/en-us/research/event/responsible-ai-an-interdisciplinary-approach-workshop/.

[288] "Xie Xing: conduct research that can withstand the test of time, develop responsible AI (谢幸：做经得起时间检验的研究，打造负责任的人工智能)."

[289] Yao et al., "From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models."

[290] Zhu et al., "PromptBench."

[291] "PAIR Lab: PKU Alignment and Interaction Research Lab," PAIR Lab: PKU Alignment and Interaction Research Lab, accessed October 12, 2023, https://pair-lab.com/.

- Key Personnel: PAIR is led by **YANG Yaodong (杨耀东)**, assistant professor at the Institute for AI at Peking University. He was previously an assistant professor at King's College London and received a PhD in Computer Science from University College London (2021). Professor Yang gave a talk on Safe Value Alignment for LLMs at the 2023 BAAI Conference.[292]
- Key Research Results:
    - OmniSafe: an open source framework providing algorithms to accelerate safe reinforcement learning. May 2023.[293]
    - Beaver: an open source library to increase LLM safety using a modified version of RLHF applying methods from Safe RL (or constrained RL) to RLHF. May 2023.[294]

RealAI

- Mission: RealAI's website states that it aims to "become a leader in safe/secure, reliable, and controllable AI infrastructure."[295]
- Key Personnel: RealAI was co-founded by Zhu Jun and **TIAN Tian (田天),** both PhD graduates from Tsinghua University. Tian serves as the startup's CEO, and Zhu serves as the company's chief scientist along with his former PhD supervisor, Zhang Bo. Besides his role at RealAI, Zhu is a professor at Tsinghua University.[296] Zhang is an academician of the Chinese Academy of Sciences and considered one of the founding figures in China's AI research due to his contributions in setting up the State Key Lab of Intelligent Technology and Systems, the first state-run AI lab in China.[297]
- Key Research Results:
    - Testing robustness of commercial multimodal large language models (MLLMs): this paper uses transfer-based attacks to test several MLLMs including Bard,

[292] "Yang Yaodong Safe Value Alignment for LLM-2023 Beijing Academy of AI Conference-AI Safety and Alignment Forum (杨耀东 Safe Value Alignment for LLM-2023北京智源大会-AI安全与对齐论坛)," Bilibili, June 11, 2023, https://www.bilibili.com/video/BV1gh411T7qS/.

[293] Jiaming Ji et al., "OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research" (arXiv, May 16, 2023), http://arxiv.org/abs/2305.09304.

[294] PKU-Alignment, "PKU Beaver: Constrained Value-Aligned LLM via Safe RLHF."

[295] "About Us: Mission and Vision (关于我们：使命愿景)," RealAI, accessed October 18, 2023, https://www.realai.ai/about?scrollId=t19.

[296] "Jun Zhu's Homepage," Department of Computer Science and Technology, Tsinghua University (清华大学计算机科学与技术系), accessed October 19, 2023, https://ml.cs.tsinghua.edu.cn/~jun/index.shtml.

[297] "Zhang Bo: A Founder of Artificial Intelligence in China (张钹：中国人工智能奠基者)," Chinese Academy of Sciences (中国科学院), August 30, 2021, https://www.cas.cn/xzfc/202108/t20210830_4803805.shtml.

Bing Chat, ERNIE bot, and GPT-4V on image description tasks. October 2023.[298]

  - ○ RealSafe: this product from RealAI supports detection of adversarial sample attacks and model backdoor attacks on various applications, including LLMs.[299]
  - ○ Blackbox backdoor detection: a method to identify malicious backdoors in deep neural networks using only query access to the model, without needing the poisoned training data or access to the white-box model. Accepted by ICCV 2021. March 2021.[300]

Shanghai AI Lab (SHLAB)

- Mission: SHLAB's website states that it seeks to conduct "strategic, innovative, and forward-looking research," in order to "achieve breakthroughs on important and fundamental theory and critical and core technologies in AI."[301] Its six research directions include AI ethics and policy, which focuses on economic, societal, ethical, legal, safety, privacy, and data governance problems triggered by AI.[302]
- Key Personnel: It is unclear which senior researchers at SHLAB are primarily working on AI safety and alignment, if any. However, SHLAB has announced recruiting for a number of positions which require knowledge in value alignment or safety and alignment problems.[303]

[298] Yinpeng Dong et al., "How Robust Is Google's Bard to Adversarial Image Attacks?" (arXiv, October 14, 2023), http://arxiv.org/abs/2309.11751.

[299] "AI Security Platform: RealSafe (人工智能安全平台RealSafe)," RealAI, accessed October 19, 2023, https://www.realai.ai/products/55.html.

[300] Yinpeng Dong et al., "Black-Box Detection of Backdoor Attacks with Limited Information and Data" (arXiv, March 24, 2021), https://doi.org/10.48550/arXiv.2103.13127.

[301] "About Us (关于我们)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), accessed October 12, 2023, https://www.shlab.org.cn/aboutus.

[302] "Research (研究方向)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), accessed October 12, 2023, https://www.shlab.org.cn/research.

[303] Eg. a "LLM Junior Researcher" with experience in value alignment, "Shanghai AI Lab Pushi Open-Source System Team | Global Recruiting (上海人工智能实验室浦视开源体系团队 | 全球招聘)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), August 17, 2023, https://www.shlab.org.cn/news/5443474; "Pushi Open Source System - Large Language Model Young Researcher (浦视开源体系-大语言模型青年研究员)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), October 11, 2023, https://www.shlab.org.cn/joinus/detail/44480b3d-fbc7-4a41-8ddb-592d46e12f82?mode=social; A Responsible AI intern working on explainability, robustness, and safety research: "JR-Large Model Trustworthy AI Algorithms Intern Researcher (JR-大模型可信AI算法见习研究员)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), October 7, 2023, https://www.shlab.org.cn/joinus/detail/f1b9cebf-3942-4a66-acd3-1613eb28173f; A junior researcher and engineer working on large model safety testing: "Large Model Safety Evaluations-Large Model Safety Youth Researcher (大模型安全评测-大模型安全青年研究员)," Shanghai Artificial Intelligence

- Key Research Results:
    - WanJuan: an openly released dataset with text, image-text, and video modalities exceeding 2TB, claiming to have value alignment because it does not have pornography, violence, and bias. September 2023.[304]

Shanghai Jiaotong University Generative AI Research Lab (GAIR)

- Mission: GAIR's stated mission is "to create cutting-edge generative AI technologies that are intelligent, ethical, transparent, and accountable, aligned with human values and serving the common good." The lab listed generative AI safety and alignment as one of six research directions in a recruiting notice in May 2023.[305]
- Key personnel: GAIR is led by **LIU Pengfei (刘鹏飞)**, an associate professor at Shanghai Jiaotong University. He is a co-founder of the AI company Inspired Cognition and received his PhD in the School of Computer Science at Fudan University (2019).[306]
- Key Research Results:
    - FacTool: a task and domain agnostic framework for detecting factual errors of texts generated by large language models. July 2023.[307]
    - ChineseFactEval: a 125 question dataset to assess factual accuracy of Chinese LLMs in seven categories. September 2023.[308]

Laboratory (上海人工智能实验室), August 2, 2023, https://www.shlab.org.cn/joinus/detail/16309e3d-ee57-43d7-88d5-f52e02f3247c?mode=social; "Large Model Safety/Security Evaluations-Large Model Safety/Security Engineer (大模型安全评测-大模型安全工程师)," Shanghai Artificial Intelligence Laboratory (上海人工智能实验室), September 14, 2023, https://www.shlab.org.cn/joinus/detail/653da0a1-9f39-4c74-9b1f-4e6b89053356?mode=campus&keyword=&zhinengId=&commitment=.

[304] Conghui He et al., "WanJuan: A Comprehensive Multimodal Dataset for Advancing English and Chinese Large Models" (arXiv, September 15, 2023), http://arxiv.org/abs/2308.10755.

[305] "GAIR: Generative Artificial Intelligence Research Lab," GAIR, accessed October 12, 2023, https://plms.ai/. Synced Review (机器之心); "(Shanghai Jiaotong University Associate Professor Liu Pengfei is recruiting NLP and Generative AI undergraduate/graduate students for his lab (上海交大副教授刘鹏飞实验室招收NLP、生成式AI方向本科生/研究生)," May 10, 2023, http://posts.careerengine.us/p/645c0068cc3d344e6a1bcc62.

[306] AlphaLab, "Q&A With Inspired Cognition Co-Founders Graham Neubig, Pengfei Liu & Yusuke Oda," *Startups & Investment* (blog), December 15, 2022, https://medium.com/startups-and-investment/q-a-with-inspired-cognition-co-founders-graham-neubig-pengfei-liu-yusuke-oda-299b433e5fce. "Pengfei Liu," Pengfei, accessed October 12, 2023, http://pfliu.com/.

[307] I.-Chun Chern et al., "FacTool: Factuality Detection in Generative AI -- A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios" (arXiv, July 26, 2023), https://doi.org/10.48550/arXiv.2307.13528. "GAIR-NLP/Factool," Python (2023; repr., Generative Artificial Intelligence Research Lab (GAIR), October 10, 2023), https://github.com/GAIR-NLP/factool.

[308] Binjie Wang, Ethan Chern, and Pengfei Liu, "ChineseFactEval: A Factuality Benchmark for Chinese LLMs," September 2023, https://gair-nlp.github.io/ChineseFactEval/.

Shanghai Jiaotong University Lab for Interpretability and Theory-Driven Deep Learning

- Mission: The lab aims to solve two problems: 1) "how to develop a theoretical system to unify different explanations of the semantics encoded in a DNN [deep neural network]" … "and unify different explanations of the DNN performance" 2) "how to extract the common mechanism behind different heuristic methods."[309]
- Key Personnel: Dr. **ZHANG Quanshi (张拳石)** leads the lab and is an associate professor at Shanghai Jiaotong University. He holds a PhD from the University of Tokyo (2014 and 2011). Zhang has released 50+ relevant papers[310] on network interpretability.
- Key Research Results:
    - Can inference logic of LLMs be disentangled into Symbolic Concepts: argues that inference score of an LLM can be disentangled into a small number of symbolic concepts that can explain prediction errors of an LLM. April 2023.[311]
    - Survey on Visual Interpretability for Deep Learning: investigates literature on interpretability of convolutional neural networks and future trends in explainable AI. February 2018, before Dr. Zhang joined Shanghai Jiaotong University.[312]

Tianjin University Natural Language Processing Laboratory (TJUNLP)

- Mission: TJUNLP's website notes that NLP is the "crowning jewel" of AI, with important value for scientific research and practical application. The four main research directions of the lab are machine translation, question answering, dialogue, and natural language generation.[313]
- Key Personnel: TJUNLP's director is Professor **XIONG Deyi (熊德意)**, who is concurrently director of the Tianjin University International Joint Research Center of Language Intelligence and Technology. Before coming to Tianjin, Dr. Xiong had been a

[309] "Lab for Interpretability and Theory-Driven Deep Learning," SJTU interpretable ML Lab, accessed October 15, 2023, https://sjtu-xai-lab.github.io/.

[310] Quanshi Zhang (张拳石), "Publications | Quanshi Zhang," accessed October 12, 2023, http://qszhang.com/index.php/publications/.

[311] Wen Shen et al., "Can the Inference Logic of Large Language Models Be Disentangled into Symbolic Concepts?" (arXiv, April 3, 2023), https://doi.org/10.48550/arXiv.2304.01083.

[312] Zhang and Zhu, "Visual Interpretability for Deep Learning."

[313] "TJUNLP Lab," TJUNLP, accessed October 12, 2023, https://tjunlp-lab.github.io/.

professor at Soochow University (2013-2018) and a research scientist at the Institute for Infocomm Research, Singapore (2007-2013).[314]

- Key Research Results:
    - A survey of Large Language Model Alignment: an extensive survey reviewing alignment, interpretability, and adversarial robustness work on LLMs, discussing sophisticated issues including inner versus outer alignment, scalable oversight, mechanistic interpretability, and deceptive alignment. September 2023.[315]
    - Watermarking conditional text generation: introducing a semantic-aware watermarking algorithm that the authors believe largely avoids performance loss for existing watermarking during conditional text generation. In collaboration with University of California, Riverside. July 2023.[316]
    - CBBQ: a Chinese bias dataset containing 100,000 questions created through human-AI collaboration, covering stereotypes and social biases in 14 dimensions relating to Chinese culture and values. June 2023.[317]

Tsinghua ConversationalAI (CoAI)

- Mission: CoAI's website states that Conversational AI means language interaction activities such as conversing, asking questions, etc., which it claims is the most challenging and most comprehensive AI technology.[318]
- Key Personnel: One leader of CoAI is **HUANG Minlie (黄民烈)**, professor in the Department of Computer Science, Institute for Artificial Intelligence, Tsinghua University. He received a PhD in Engineering from Tsinghua University (2006).[319] Professor Huang gave a speech on LLM safety at the BAAI Conference in 2023, discussing how he used red teaming to improve model safety, and presenting work on large model safety evaluations.
- Key Research Results:

[314] "Home," Deyi Xiong (熊德意), accessed October 12, 2023, https://dyxiong.github.io/index.html.

[315] Shen et al., "Large Language Model Alignment."

[316] Yu Fu, Deyi Xiong, and Yue Dong, "Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy" (arXiv, July 25, 2023), https://doi.org/10.48550/arXiv.2307.13808.

[317] Yufei Huang and Deyi Xiong, "CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models" (arXiv, June 28, 2023), https://doi.org/10.48550/arXiv.2306.16244.

[318] 交互式人工智能课题组, "CoAI (清华大学交互式人工智能课题组)," CoAI, accessed October 12, 2023, https://coai.cs.tsinghua.edu.cn/.

[319] "Huang Minlie (黄民烈)," Department of Computer Science and Technology, Tsinghua University (清华大学计算机科学与技术系), March 31, 2021, https://www.cs.tsinghua.edu.cn/info/1121/5620.htm.

  - MoralDial: a framework using conversations between simulated users and the dialogue system on explaining, revising, and inferring moral dialogues to teach models morality. Also co-led by Huawei Noah's Ark Lab. May 2023.[320]
  - Safety Assessment of Chinese Large Language Models: a benchmark assessing Chinese LLM safety in eight typical safety scenarios and six types of instruction attacks. April 2023.[321]

Tsinghua Foundation Model Research Center

- Mission: The center seeks to "conduct research and pursue applications of AI foundation models." The four areas of focus are: foundation model theory and core technology; overlap between foundation models and other academic disciplines; exchanges involving industry, academia, and researchers; and developing talents relating to foundation models.[322]
- Key Personnel: The director of the center is **TANG Jie (唐杰)**. Dr. Tang is also part of the Tsinghua Knowledge Engineering Group (KEG), which has released some of China's most powerful bilingual LLMs, GLM-130B and ChatGLM. His research is focused on "**artificial general intelligence** with a mission toward **teaching machines to think like humans** [emphasis original]."[323] He is a fellow of the Association for Computing Machinery (ACM), Association for the Advancement of Artificial Intelligence (AAAI), and Institute of Electrical and Electronics Engineers (IEEE), and received a PhD in engineering from Tsinghua University (2006). A deputy director of the center is Huang Minlie, who was described earlier in the section on CoAI.
- Key Research Results:
  - The center was only founded in August 2023, and thus does not have many results published. In September 2023, they announced publication of SafetyBench: a safety benchmark consisting of 11,435 multiple choice questions across seven categories of safety concerns — offensiveness;

[320] Sun et al., "MoralDial."

[321] Sun et al., "Safety Assessment of Chinese Large Language Models."

[322] Tsinghua University Foundation Models (THU基础模型), "'About Tsinghua University Foundation Models Research Center' (【清华大学基础模型研究中心简介】)," Weixin Official Accounts Platform, September 25, 2023, http://mp.weixin.qq.com/s?__biz=MzkwMzU1MDMzOQ==&mid=2247484067&idx=1&sn=49cc17fba9ab77814b0adf680a917685&chksm=c095c64ff7e24f59b3f2fd3cb475ddf5dc2222cf32e42f2135e6b5bed421779ec43162d8141b#rd.

[323] "Jie Tang (Tang, Jie) 唐杰," KEG, accessed October 19 2023, https://keg.cs.tsinghua.edu.cn/jietang/.

unfairness and bias; physical health; mental health; illegal activities; ethics and morality; and privacy and property.[324] The benchmark was developed in collaboration with Tsinghua CoAI and KEG.

[324] Tsinghua University Foundation Models (THU基础模型), "SafetyBench: Evaluating the Security of Large Language Models through Multiple Choice Questions (SafetyBench：通过单选题评估大型语言模型安全性)," Weixin Official Accounts Platform, September 20, 2023, http://mp.weixin.qq.com/s?__biz=MzkwMzU1MDMzOQ==&mid=2247484013&idx=1&sn=05dfb8ea209ae5063d00037e25bb8cbd&chksm=c095c681f7e24f9745ee230d5a4c369a65aad9f7d11b3ed9a804fcc6beed25442ddc964f665b#rd. Zhang et al., "SafetyBench."

Appendix D. Notable AI Safety Papers by Chinese Research Groups

General

- Wang, Jindong, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, et al. "On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective." arXiv, August 29, 2023. http://arxiv.org/abs/2302.12095.
- Yao, Jing, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. "From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models." arXiv, September 3, 2023. http://arxiv.org/abs/2308.12014.

Specification

- Deng, Jiawen, Hao Sun, Zhexin Zhang, Jiale Cheng, and Minlie Huang. "Recent Advances towards Safe, Responsible, and Moral Dialogue Systems: A Survey." arXiv, March 6, 2023. http://arxiv.org/abs/2302.09270.
- Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." arXiv, August 10, 2023. http://arxiv.org/abs/2308.05374.
- PKU-Alignment. "PKU Beaver: Constrained Value-Aligned LLM via Safe RLHF." Accessed October 10, 2023. https://pku-beaver.github.io/.
- PKU-Alignment. "BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset." Accessed October 10, 2023. https://sites.google.com/view/pku-beavertails/home.
- Sun, Hao, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. "On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark." arXiv, April 4, 2022. https://doi.org/10.48550/arXiv.2110.08466.
- Sun, Hao, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. "MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions." arXiv, May 26, 2023. http://arxiv.org/abs/2212.10720.

- Yao, Jing, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. "From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models." arXiv, September 3, 2023. http://arxiv.org/abs/2308.12014.
- Yuan, Luyao, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. "In Situ Bidirectional Human-Robot Value Alignment." *Science Robotics* 7, no. 68 (July 13, 2022): eabm4183. https://doi.org/10.1126/scirobotics.abm4183.
- Zhang, Zhaowei, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. "Heterogeneous Value Evaluation for Large Language Models." arXiv, June 1, 2023. http://arxiv.org/abs/2305.17147.
- Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. "SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions." arXiv, September 13, 2023. http://arxiv.org/abs/2309.07045.

Robustness

- Chen, Xuanting, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. "How Robust Is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks." arXiv, March 1, 2023. https://doi.org/10.48550/arXiv.2303.00293.
- Dong, Yinpeng, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. "How Robust Is Google's Bard to Adversarial Image Attacks?" arXiv, October 14, 2023. http://arxiv.org/abs/2309.11751.
- Gui, Tao, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, et al. "TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing." arXiv, May 5, 2021. http://arxiv.org/abs/2103.11441.
- Li, Linyang, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. "BERT-ATTACK: Adversarial Attack Against BERT Using BERT." arXiv, October 1, 2020. http://arxiv.org/abs/2004.09984.
- Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. "Backdoor Learning: A Survey." arXiv, February 16, 2022. http://arxiv.org/abs/2007.08745.
- Wang, Jindong, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, et al. "On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective." arXiv, August 29, 2023. http://arxiv.org/abs/2302.12095.

- Wu, Baoyuan, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. "BackdoorBench: A Comprehensive Benchmark of Backdoor Learning," 2022. https://openreview.net/forum?id=31_U7n18gM7.
- Wu, Baoyuan, Li Liu, Zihao Zhu, Qingshan Liu, Zhaofeng He, and Siwei Lyu. "Adversarial Machine Learning: A Systematic Survey of Backdoor Attack, Weight Attack and Adversarial Example." arXiv, February 18, 2023. http://arxiv.org/abs/2302.09457.
- Zhu, Kaijie, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, et al. "PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts." arXiv, August 24, 2023. http://arxiv.org/abs/2306.04528.

Assurance

- "BAAI/COIG-PC · Datasets at Hugging Face," October 5, 2023. https://huggingface.co/datasets/BAAI/COIG-PC.
- Chern, I.-Chun, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. "FacTool: Factuality Detection in Generative AI -- A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios." arXiv, July 26, 2023. https://doi.org/10.48550/arXiv.2307.13528.
- Deng, Jiawen, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. "COLD: A Benchmark for Chinese Offensive Language Detection." arXiv, October 19, 2022. http://arxiv.org/abs/2201.06025.
- Huang, Yufei, and Deyi Xiong. "CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models." arXiv, June 28, 2023. https://doi.org/10.48550/arXiv.2306.16244.
- Sun, Hao, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. "On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark." arXiv, April 4, 2022. https://doi.org/10.48550/arXiv.2110.08466.
- Sun, Hao, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. "Safety Assessment of Chinese Large Language Models." arXiv, April 20, 2023. https://doi.org/10.48550/arXiv.2304.10436.
- Wang, Binjie, Ethan Chern, and Pengfei Liu. "ChineseFactEval: A Factuality Benchmark for Chinese LLMs," 2023. https://gair-nlp.github.io/ChineseFactEval/

- Xu, Guohai, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, et al. "CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility." arXiv, July 18, 2023. http://arxiv.org/abs/2307.09705.
- Yu, Yaodong, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. "White-Box Transformers via Sparse Rate Reduction." arXiv, June 1, 2023. https://doi.org/10.48550/arXiv.2306.01129.
- Zhang, Quanshi, and Song-Chun Zhu. "Visual Interpretability for Deep Learning: A Survey." arXiv, February 7, 2018. http://arxiv.org/abs/1802.00614.
- Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. "SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions." arXiv, September 13, 2023. http://arxiv.org/abs/2309.07045.
- Zhou, Bolei, David Bau, Aude Oliva, and Antonio Torralba. "Comparing the Interpretability of Deep Networks via Network Dissection." In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 243–52. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-28954-6_12.
- Zhou, Jingyan, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. "Towards Identifying Social Bias in Dialog Systems: Frame, Datasets, and Benchmarks." arXiv, October 28, 2022. http://arxiv.org/abs/2202.08011.

Appendix E. List of Lab Safety/Ethics Values

Institution	Release date and context	Safety/Ethics values
Alibaba	Unclear; Alibaba established the Alibaba Artificial Intelligence Governance Laboratory (AAIG) in 2021.[325] The AAIG's three values are "usability," "reliability," and "credibility," and it claims to have published over 100 papers in international conferences or periodicals.[326]	Alibaba established an S&T ethics governance committee in September 2022.[327] The committee announced six values: people-centeredness, inclusivity and honesty, safety and reliability, privacy protection, trustworthiness and controllability, and openness and co-governance.
Ant Group	March 2, 2023, stated by General Counsel **ZHOU Zhifeng (周志峰)** when Ant Group announced the creation of an S&T ethics advisory committee in March 2023. This supplemented an existing technology ethics committee in-house.[328] In September 2023, Ant also announced a plan for developing ethics for large models.[329]	General Counsel Zhou Zhifeng stated that the four values for the ethics committee are "equality, respect, trust, and responsibility."[330]
BAAI, Peking University,	May 25, 2019, as the "Beijing AI Principles."[331] This document is a broad statement about what the signatories	The document provides seven R&D values, three use values, and five governance values. R&D should control risks around "maturity,

[325] "Alibaba Announces the Establishment of Artificial Intelligence Governance and Sustainable Development Laboratory (阿里巴巴宣布成立人工智能治理与可持续发展实验室)," Synced Review (机器之心), July 13, 2021, https://www.jiqizhixin.com/articles/2021-07-13-13.

[326] "Alibaba Artificial Intelligence Governance Research Center (阿里巴巴人工智能治理与可持续发展实验室) (AAIG)," Alibaba, accessed October 13, 2023, https://s.alibaba.com/cn/aaig/academic-committee.

[327] "Alibaba Group CTO Cheng Li: the science and technology ethics governance committee should be the 'gatekeeper' of technological innovation (阿里集团CTO程立：科技伦理治理委员会要做技术创新的'守门人')."

[328] "Ant Group Has Establishes a Science and Technology Ethics Advisory Committee, Consisting of 7 Experts in the Fields of Science and Technology and Social Sciences (蚂蚁集团成立科技伦理顾问委员会，由7位科技及社会科学领域专家构成)," Xinhua, March 6, 2023, https://www.xinhuanet.com/tech/20230306/24e6bf97479d4eae8be12bf76d9b812b/c.html. "Science and Technology Innovation (科技创新)," Ant Group (蚂蚁集团), accessed October 13, 2023, https://www.antgroup.com/esg/innovation.

[329] Yan Yu (喻琰), "The Bund Conference | Ant Group Launches the Ethical Co-Construction Plan for Large Models, Giving Legal Exams to the Large Model (外滩大会｜蚂蚁集团启动大模型伦理共建计划，给大模型出法律考题)," The Paper (澎湃), September 9, 2023, https://m.thepaper.cn/newsDetail_forward_24548122.

[330] "Ant Group Has Establishes a Science and Technology Ethics Advisory Committee, Consisting of 7 Experts in the Fields of Science and Technology and Social Sciences (蚂蚁集团成立科技伦理顾问委员会，由7位科技及社会科学领域专家构成)," Xinhua, March 6, 2023, https://www.xinhuanet.com/tech/20230306/24e6bf97479d4eae8be12bf76d9b812b/c.html.

[331] "Beijing Artificial Intelligence Principles," International Research Center for AI Ethics and Governance, accessed October 13, 2023, https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/. "Beijing AI Principles (人工智能北京共识)," Beijing Academy of AI (北京智源研究院), accessed October 13, 2023, https://www.baai.ac.cn/portal/article/index/type/center_result/id/110.html. "The 'Beijing AI Principles' Holds up 'Harmonious and Optimized Coexistence' (人工智能'北京共识'提出'和谐与优化共生')," State Council, May 26, 2019, https://www.gov.cn/xinwen/2019-05/26/content_5394829.htm.

Tsinghua University, CAS, and the New Generation AI Industry Technology Innovation Strategic Alliance (joint release)	believe AI R&D, use, and development should look like.	robustness, reliability, and controllability." It should also consider ethical, legal, and social impacts and risks. On governance, the document called for encouraging research on the potential risks of AGI and superintelligence, thus ensuring that AI "will always be beneficial to society and nature in the future."
Baidu	November 25, 2019, published on the website for Baidu's AI platform and repeated since by CEO Robin Li.[332]	The four values listed are safety and controllability; promoting more equal technological benefits to humanity; teaching humans and helping humans grow; and bringing more freedom and possibility to humans. The statement of values also notes that AI should not surpass or replace humans and that "safety and controllability" is the highest value.
Huawei	Unclear; drawn from a page on AI security on Huawei's Trust Center website and a 2018 White Paper on AI Security.[333]	Rather than outlining values, Huawei listed five major security challenges of AI systems: hardware and software security; data integrity; model confidentiality; model robustness; data privacy. These focus on cyber and data security issues for AI, rather than risks from frontier AI systems.
Megvii	July 17, 2019, in Megvii's Artificial Intelligence Application Guidelines.[334]	The document lists six values: legitimacy; human oversight; reliability and safety of technology; fairness and diversity; accountability and timely correction; and data security and protection of personal information. The "human oversight" principle is about ensuring AI does not exceed bounds of human control. The "reliability and safety" principle is about the ability to defend against attacks and sufficiently testing AI.

[332] "Self-Discipline Principle Statement (自律性原则申明)," Baidu, November 25, 2019, https://ai.baidu.com/ai-doc/REFERENCE/xk3dwjgfe. "Robin Li: Adhere to Four Principles for AI Ethics, Create Responsible, Sustainable AI (李彦宏：坚守人工智能伦理四原则 打造负责、可持续AI)," Sina, December 27, 2021, https://finance.sina.com.cn/tech/2021-12-27/doc-ikyakumx6675750.shtml.

[333] "AI Security (AI安全)," Huawei, accessed October 13, 2023, https://www.huawei.com/cn/trust-center/ai-section; "AI Security White Paper (AI安全白皮书)," Huawei, accessed October 19 2023, https://www-file.huawei.com/-/media/corporate/pdf/cyber-security/ai-security-white-paper-cn.pdf.

[334] "Megvii's 'AI Application Guidelines' Full Document Is Published, Advocating for Beneficial AI Technology (旷视《人工智能应用准则》全文公布 提倡善用AI技术)," 36kr, July 17, 2019, https://36kr.com/p/1724031811585.

SenseTime	January 2020, from SenseTime's 2022 report on AI ethics and governance practices and report on AI ethics and governance frontier and practice.[335]	SenseTime's view of AI ethics is "balanced development," and its ethics values are "human-centeredness, controllable technology, and sustainable development." SenseTime sent up an AI Ethics and Governance Committee in January 2020. According to SenseTime, this committee consists of a secretariat, implementing working group, and advisory committee, with both internal and external members.[336]
Tencent Research Institute	December 3, 2018, stated in a speech by Dean Jason Si, building upon Tencent CEO Pony Ma's speech at the 2018 World AI Conference.[337]	Tencent's ethical framework for AI governance is summarized by the ARCC concept: availability, reliability, comprehensibility, and controllability. AI should be available to the masses; safe and reliable against accidents; more comprehensible; and with human beings in charge always.

[335] "'Balanced Development' AI Governance White Paper ('平衡发展'的人工智能治理白皮书)," SenseTime (商汤), September 2022, https://oss.sensetime.com/20221019/d6fb83bdf9e069c70890d7d85214b23e/%E5%B9%B3%E8%A1%A1%E5%8F%91%E5%B1%95%E7%9A%84AI%E6%B2%BB%E7%90%86%E7%99%BD%E7%9A%AE%E4%B9%A6.pdf; "SenseTime S&T AI Ethics and Governance Frontier Practices (商汤科技人工智能伦理与治理前沿实践)," SenseTime (商汤), 2022, https://oss.sensetime.com/20221123/e073a90ac65fb9dadc03b7e33a94d414/%E5%95%86%E6%B1%A4%E7%A7%91%E6%8A%80AI%E4%BC%A6%E7%90%86%E4%B8%8E%E6%B2%BB%E7%90%86%E5%89%8D%E6%B2%BF%E5%AE%9E%E8%B7%B5-PDF%E6%8A%A5%E5%91%8A.pdf.

[336] "SenseTime AI Ethics and Governance Committee Introduction (商汤科技人工智能伦理与治理委员会简介)," SenseTime (商汤), October 14, 2022, https://www.sensetime.com/cn/ethics-detail/56847?categoryId=32429.

[337] "Build an ethical "ark" to make artificial intelligence knowable, controllable, available and reliable (打造伦理"方舟"，让人工智能可知、可控、可用、可靠)," World Economic Forum, December 27, 2018 https://cn.weforum.org/agenda/2018/12/b9b7b1dc-5021-4f7d-965b-f4cf3bb4cb3b; Jason Si, "These Rules Could Save Humanity from the Threat of Rogue AI," World Economic Forum, May 8, 2019, https://www.weforum.org/agenda/2019/05/these-rules-could-save-humanity-from-the-threat-of-rogue-ai/. "Shanghai 2018 World Artificial Intelligence Conference Kicks off, China's AI Biggest Players Surfaced," China Knowledge, September 18, 2018, https://www.chinaknowledge.com/News/DetailNews?id=80279.

Appendix F. Key Surveys on AI in China

Pollster	Sample time	Sample breakdown	Key results
Communist Youth Daily, the official newspaper of the Communist Youth League[338]	June 2017	2006 individuals sampled, with 21% <27 years old, 51% 27-37 years old, 21% 37-47 years old, and 6% over 47 years old.	Respondents identified overreliance on AI (63%) and privacy (53%) as the main risks associated with AI.
Center for International Security and Strategy (CISS), a think tank at Tsinghua University[339]	Unclear, the report was published in 2019	1,491 individuals sampled, all between ages of 14 and 28. 13% were computer science students, and 72% were students not in computer science. 78% were in a standard four year college or higher level of education.	Respondents were most worried about risks of AI in employment (52%) and privacy and ethics (43%). Respondents wanted to prevent and regulate AI-related risks through passing (unspecified) laws (68%), promoting vocational education to stimulate employment transformation (~60%), and to a lesser degree increasing public awareness (~36%). There was minimal support for reducing R&D, maintaining the status quo, or taxing AI companies (<10% each).
Cheetah Mobile, a Chinese mobile internet company[340]	Unclear, the report was published in September 2019	3625 individuals sampled, 45% were 25 years old or younger, 6% were aged 60+, and 14% were aged 51-60. 17% had a Bachelor's degree or higher. 40% were in a fourth-tier city and 17% were in a third-tier city.	Respondents' "feelings" towards AI development were: 61% expectant, 56% excited, 9% worried (担忧), 5% skeptical (质疑), and 3% anxious (焦虑).
Center for Long-term Artificial Intelligence	March to July 2021	1032 individuals sampled, primarily young and middle-aged AI-related students and scholars, as	86% of respondents thought that the impact of "continuous research and application of large AI models on society" is

[338] "77.8% of Respondents Are Optimistic about the Development of Artificial Intelligence in China (77.8%受访者看好我国人工智能的发展)," Communist Youth Daily (中国青年报), June 13, 2017, https://zqb.cyol.com/html/2017-06/13/nw.D110000zgqnb_20170613_1-07.htm.

[339] 战略与安全研究中心. "Pre-Research Report: Risks and Governance of AI from the Perspective of Chinese Youth" (Center for International Security and Strategy Tsinghua University (清华大学战略与安全研究中心), 2019), https://ciss.tsinghua.edu.cn/upload_files/atta/1589025522890_83.pdf.

[340] Cheetah User Research Center (猎豹用户研究中心), "Cheetah Announcement | AI in the Eyes of Ordinary People: A Research Report on Public Recognition, Feelings, and Attitudes towards AI (豹告 | 普通人眼中的AI：大众AI认知、感受、态度调研报告)," Weixin Official Accounts Platform, September 9, 2019, http://mp.weixin.qq.com/s?__biz=MzU3NjI0MzgxOQ==&mid=2247489828&idx=1&sn=2efbbed9c94c8fbb98e5fcbc2700417c&chksm=fd178820ca6001361dc29bacfd925439e06505b3efdcd8475d8a738c21028ac0a5a628a21d2f#rd.

(CLAI), led by Professor Zeng Yi from the Chinese Academy of Sciences[341]		well as 63 experts invited to participate through scientific industry associations. Overall, 15% worked in AI R&D, 14% were AI researchers or scholars, 27% were in other natural science fields, 5% were humanities or social science scholars, and 38% were from other professions.	positive. 76-82% of respondents thought that "Strong AI can be achieved." 64-85% of respondents thought that "Strong AI should be developed." 65-73% thought that "Strong AI can coexist harmoniously" with humans. 60-73% of respondents thought that "Strong AI poses existential threats to humans." A subset of 63 experts with honors from several Chinese AI or computer science professional associations or institutions believed that Strong AI poses existential risks (72%), but supported developing Strong AI (79%) and thought that Strong AI can coexist harmoniously with humans (69%).
Ipsos, a French market research and survey company, on behalf of the World Economic Forum[342]	November 19 to December 3, 2021	Approximately 1,000 individuals sampled in China (mainland). The sample in China (mainland) is "more urban, more educated, and/or more affluent than the general population" and reflects the views of the "more 'connected' segment" of the population.	Respondents in China thought that AI products and services "have more benefits than drawbacks" 78% of the time. This was highest of all 28 countries surveyed and much higher than the overall average of 52%. Chinese respondents also had the highest level of trust in companies that use AI and a low level of nervousness about using AI products (4th lowest among 28 countries).
Center for Long-term Artificial Intelligence (CLAI), led by Professor Zeng Yi from the Chinese	March to April 2023	566 individuals sampled, with 44% in AI, 21% in other computer science and engineering fields, and 16% in psychology or cognitive science. 92% knew about "giant AI models" and 85% often or occasionally used	33% of respondents supported a six-month pause on training AI systems more advanced than GPT-4; 31% did not support it; and 30% thought a pause would be "useless." 91% of respondents supported

[341] Yi Zeng (曾毅) and Kang Sun, "Whether We Can and Should Develop Strong AI: A Survey in China," Center for Long-term Artificial Intelligence, Center for Long-term Artificial Intelligence, March 12, 2023, https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence.

[342] Nicolas Boyon, "Opinions about AI Vary Depending on Countries' Level of Economic Development," Ipsos, January 5, 2022, https://www.ipsos.com/en/global-opinions-about-ai-january-2022; "Global Opinions and Expectations About Artificial Intelligence: A Global Advisor Survey" (Ipsos, January 2022), https://www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Global-opinions-and-expectations-about-AI-2022.pdf.

Academy of Sciences[343]		products or services relating to giant AI models.	implementation of an “ethics, safety and governance framework for every large AI model used in social services.”

[343] Yi Zeng (曾毅) et al., “Voices from China on ‘Pause Giant AI Experiments: An Open Letter,’” Center for Long-term Artificial Intelligence, April 4, 2023, https://long-term-ai.center/research/f/voices-from-china-on-pause-giant-ai-experiments-an-open-letter.

Appendix G. Key Acronyms and Translations

AI	Artificial Intelligence	人工智能
AIDP	New Generation Artificial Intelligence Development Plan	新一代人工智能发展规划
AGI	Artificial General Intelligence	通用人工智能
AIIA	Artificial Intelligence Industry Alliance	人工智能产业发展联盟
BAAI	Beijing Academy of Artificial Intelligence	北京智源人工智能研究院
CAC	Cyberspace Administration of China	网信办
CAICT	China Academy of Information and Communications Technology	中国信息通信研究院
CAIS	Center for AI Safety	人工智能安全中心
CCW (or CCWC)	Convention on Certain Conventional Weapons	特定常规武器公约
CESI	China Electronics Standardization Institute	中国电子技术标准化研究院
CISS	Tsinghua University's Center for International Security and Strategy	清华大学国际安全与战略研究中心
CLAI	Center for Long-term Artificial Intelligence	远期人工智能研究中心
CPC	Communist Party of China	中国共产党
CUHK	Chinese University of Hong Kong	香港中文大学
EU	European Union	欧盟
FLI	Future of Life Institute	生命未来研究所
GAIR	Generative Artificial Intelligence Research Lab at Shanghai Jiaotong University	上海交通大学清源研究院生成式人工智能研究组
HKU	The University of Hong Kong	香港大学
HKUST	Hong Kong University of Science and Technology	香港科技大学
IAEA	International Atomic Energy Agency	国际原子能机构

ICAO	International Civil Aviation Organization	国际民用航空组织
IEEE	Institute of Electrical and Electronics Engineers	电气与电子工程师协会
ISO	International Organization for Standardization	国际标准化组织
IPCC	Intergovernmental Panel on Climate Change	联合国政府间气候变化专门委员会
LAWs	Lethal Autonomous Weapons	致命自主武器
LLM	Large Language Model	大语言模型
MIIT	Ministry of Industry and Information Technology	工信部
MOFA	Ministry of Foreign Affairs	外交部
MOST	Ministry of Science and Technology	科技部
OECD	Organisation for Economic Co-operation and Development	经济合作与发展组织
PKU	Peking University	北京大学
R&D	Research and Development	研究与发展
RLHF	Reinforcement Learning from Human Feedback	人类反馈强化学习
SAC	Standardization Administration of China	中国标准化管理委员会
SHLAB	Shanghai Artificial Intelligence Laboratory	上海人工智能实验室
S&T	Science and Technology	科学与技术
UK	United Kingdom	英国
UN	United Nations	联合国
UNSC	United Nations Security Council	联合国安全理事会
US	United States	美国
WAIC	World Artificial Intelligence Conference	世界人工智能大会
ZGC	Zhongguancun Forum	中关村论坛

References

Agüera Y Arcas, Blaise and Peter Norvig. “Artificial General Intelligence Is Already Here.” *Noema Magazine*, October 10, 2023. https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

AI for Good. “AI For Good Global Summit 2023.” Accessed October 11, 2023. https://aiforgood.itu.int/summit23/.

Alibaba. “Alibaba Artificial Intelligence Governance Research Center (阿里巴巴人工智能治理与可持续发展实验室) (AAIG).” Accessed October 13, 2023. https://s.alibaba.com/cn/aaig/academic-committee.

Alibaba DAMO Academy. “About - DAMO Academy.” Accessed October 12, 2023. https://damo.alibaba.com/about/.

AlphaLab. “Q&A With Inspired Cognition Co-Founders Graham Neubig, Pengfei Liu & Yusuke Oda.” *Startups & Investment* (blog), December 15, 2022. https://medium.com/startups-and-investment/q-a-with-inspired-cognition-co-founders-graham-neubig-pengfei-liu-yusuke-oda-299b433e5fce.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. “Concrete Problems in AI Safety.” arXiv, July 25, 2016. https://doi.org/10.48550/arXiv.1606.06565.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, et al. “Frontier AI Regulation: Managing Emerging Risks to Public Safety.” arXiv, September 4, 2023. https://doi.org/10.48550/arXiv.2307.03718.

Andrew Yao (姚期智). “Andrew YAO.” Translated by Concordia AI. Chinese Perspectives on AI. Accessed October 12, 2023. https://chineseperspectives.ai/Andrew-YAO.

Ant Group (蚂蚁集团). “Science and Technology Innovation (科技创新).” Accessed October 13, 2023. https://www.antgroup.com/esg/innovation.

António Guterres. “Secretary-General’s Remarks to the Security Council on Artificial Intelligence.” United Nations Secretary-General, July 18, 2023. https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). “About AIIA (关于联盟).” Accessed October 15, 2023. http://www.aiiaorg.cn/index.php?m=alliance&c=index&a=structure&s=2#zjjg.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). “AIIA Overall Group 2020 Report on the Rationale behind Our Work (AIIA总体组2020年工作思路汇报),” May 7, 2020. http://www.aiiaorg.cn/index.php?m=content&c=index&a=show&catid=2&id=263.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). “Groups (工作组&推进组&委员会).” Accessed October 16, 2023. http://www.aiiaorg.cn/index.php?m=alliance&c=index&a=workgroups&mgroup=3.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). “Notice on Preparing to Establish the AIIA ‘Artificial Intelligence Value Alignment Partnership Plan’ and Soliciting the First Batch of Member Units (关于筹备成立AIIA‘人工智能价值对齐伙伴计划’并征集首批成员单位的通知).” Weixin Official Accounts Platform, October 8, 2023. https://mp.weixin.qq.com/s/rzw-zTB2bO34Aeun6oHZ2g.

———. “Notice on the Preparation for the Establishment of the AIIA Safety/Security Governance Committee: Solicitation for the First Batch of Member Units Simultaneously Starts (关于筹备成立AIIA安全治理委员会的通知首批成员单位同步开始征集).” Weixin Official Accounts Platform, September 27, 2023. http://mp.weixin.qq.com/s?__biz=MzU0MTEwNjg1OA==&mid=2247501263&idx=1&sn=b7dfeb79135ef1f27faaee7137b320f5&chksm=fb2c720acc5bfb1cec2752c16933bbaf92837a432bec282243f658941608f1b68f125bcfc7ed#rd.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). "Services (联盟服务)," March 29, 2019. http://aiiaorg.cn/index.php?m=content&c=index&a=show&catid=34&id=58.

Artificial Intelligence Industry Alliance (AIIA) (人工智能产业发展联盟). "The Artificial Intelligence Industry Alliance Signed the 'AI Industry Self-Governance Convention' Proposal (人工智能产业发展联盟签署《人工智能行业自律公约》的倡议)," August 8, 2019. http://www.aiiaorg.cn/index.php?m=content&c=index&a=show&catid=3&id=49.

———. "Trustworthy AI Technology Hotspot | Large Models Continually Release Technological Dividends, and an Evaluation System for the Extremely Large Model Industry Is Formally Released (可信AI技术热点丨大模型持续释放技术红利，产业级大模型评估体系正式发布)." Weixin Official Accounts Platform, June 27, 2022. http://mp.weixin.qq.com/s?__biz=MzU0MTEwNjg1OA==&mid=2247499125&idx=2&sn=fc677dcdd56cc78b59563798bfedc2c7&chksm=fb2c4ab0cc5bc3a687deebc43d07a53b3e0ac79829fc77a4b8cbd4630824c622c2191005dbf2#rd.

Astha Rajvanshi. "Rishi Sunak Wants the U.K. to Be a Key Player in Global AI Regulation." Time, June 14, 2023. https://time.com/6287253/uk-rishi-sunak-ai-regulation/.

"BAAI/COIG-PC · Datasets at Hugging Face," October 5, 2023. https://huggingface.co/datasets/BAAI/COIG-PC.

Baichuan Inc. "Baichuan 2: Open Large-Scale Language Models." Accessed October 19 2023. https://cdn.baichuan-ai.com/paper/Baichuan2-technical-report.pdf.

Baidu. "Self-Discipline Principle Statement (自律性原则申明)," November 25, 2019. https://ai.baidu.com/ai-doc/REFERENCE/xk3dwjgfe.

Baidu Baike (百度百科). "2018 World Artificial Intelligence Conference (2018世界人工智能大会)." Accessed October 12, 2023. https://baike.baidu.com/item/2018%E4%B8%96%E7%95%8C%E4%BA%BA%E5

%B7%A5%E6%99%BA%E8%83%BD%E5%A4%A7%E4%BC%9A/22705586?fr=ge_ala#3.

Baidu Baike (百度百科). "Directly-Administered Municipalities (直辖市)." Accessed October 17, 2023. https://baike.baidu.com/item/%E7%9B%B4%E8%BE%96%E5%B8%82/725471.

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Network Dissection: Quantifying Interpretability of Deep Visual Representations." In *Computer Vision and Pattern Recognition*, 2017.

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. "Understanding the Role of Individual Units in a Deep Neural Network." *Proceedings of the National Academy of Sciences*, 2020. https://doi.org/10.1073/pnas.1907375117.

Beijing Academy of AI (北京智源研究院). "2021 Beijing Academy of AI Conference | AI Ethics, Governance, and Sustainable Development Forum - Morning (2021 北京智源大会 | 人工智能伦理、治理与可持续发展论坛 - 上午)," May 28, 2021. https://hub.baai.ac.cn/view/8306.

Beijing Academy of AI (北京智源研究院). "2021 Beijing Academy of AI Conference | Pre-Trained Models Forum (2021北京智源大会 | 预训练模型论坛)," May 28, 2021. https://hub.baai.ac.cn/view/8296.

Beijing Academy of AI (北京智源研究院). "2023 北京智源大会." Accessed October 19, 2023. https://2023.baai.ac.cn/.

Beijing Academy of AI (北京智源研究院). "Beijing AI Principles (人工智能北京共识)." Accessed October 13, 2023. https://www.baai.ac.cn/portal/article/index/type/center_result/id/110.html.

Beijing Academy of AI (北京智源研究院). "Special Forum on Artificial Intelligence Ethics, Governance, and Sustainable Development (人工智能伦理、治理与可持续发展专题论坛)," July 30, 2020. https://hub.baai.ac.cn/view/1681.

Beijing Economics and Informatization Bureau (北京市经济和信息化局). "Beijing Artificial General Intelligence Industry Innovation Partnership Plan (北京市通

用人工智能产业创新伙伴计划),” May 19, 2023. https://www.beijing.gov.cn/zhengce/zhengcefagui/202305/t20230524_3111706.html.

Bilibili. “Yang Yaodong Safe Value Alignment for LLM-2023 Beijing Academy of AI Conference-AI Safety and Alignment Forum (杨耀东 Safe Value Alignment for LLM-2023北京智源大会-AI安全与对齐论坛),” June 11, 2023. https://www.bilibili.com/video/BV1gh411T7qS/.

Bo Zhang (张钹). “Bo ZHANG.” Translated by Concordia AI. Chinese Perspectives on AI. Accessed October 12, 2023. https://chineseperspectives.ai/Bo-ZHANG.

Bolei Zhou, “Interpreting Deep Generative Models for Interactive AI Content Creation by Bolei Zhou (CUHK),” 2021. https://www.youtube.com/watch?v=PtRU2B6Iml4.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. “Discovering Latent Knowledge in Language Models Without Supervision.” arXiv, December 7, 2022. http://arxiv.org/abs/2212.03827.

Caffarena, Anna. “Why China’s Understanding of Multilateralism Matters for Europe,” April 2022. http://www.eurics.eu/upload/document/20220427040433_why-chinas-understanding-multilateralism-matters-for-europe-eurics-2022.pdf.

Carroll, Micah, Alan Chan, Henry Ashton, and David Krueger. “Characterizing Manipulation from AI Systems.” arXiv, March 16, 2023. https://doi.org/10.48550/arXiv.2303.09387.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, et al. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.” arXiv, September 11, 2023. https://doi.org/10.48550/arXiv.2307.15217.

Center for AI Safety. “The Landscape of US AI Legislation.” AI Safety Newsletter, September 19, 2023. https://newsletter.safe.ai/p/the-landscape-of-us-ai-legislation.

Center for AI Safety (CAIS). “Statement on AI Risk.” Accessed October 12, 2023. https://www.safe.ai/statement-on-ai-risk#open-letter.

Center for Security and Emerging Technology. “Translation: Artificial Intelligence Security Standardization White Paper.” Accessed October 11, 2023. https://cset.georgetown.edu/publication/artificial-intelligence-security-standardization-white-paper-2019-edition/.

Center for Security and Emerging Technology. “Translation: Artificial Intelligence Standardization White Paper,” May 12, 2020. https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper/.

Center for Security and Emerging Technology. “Translation: Artificial Intelligence Standardization White Paper (2021 Edition),” October 21, 2021. https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper-2021-edition/.

Center for Security and Emerging Technology. “Translation: Ethical Norms for New Generation Artificial Intelligence Released,” October 21, 2021. https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.

Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, et al. “Harms from Increasingly Agentic Algorithmic Systems.” In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–66, 2023. https://doi.org/10.1145/3593013.3594033.

Chan, Kwan Ho Ryan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. “ReduNet: A White-Box Deep Network from the Principle of Maximizing Rate Reduction.” *Journal of Machine Learning Research* 23, no. 114 (2022): 1–103.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, et al. “A Survey on Evaluation of Large Language Models.” arXiv, August 28, 2023. https://doi.org/10.48550/arXiv.2307.03109.

Cheetah User Research Center (猎豹用户研究中心). “Cheetah Announcement | AI in the Eyes of Ordinary People: A Research Report on Public Recognition, Feelings, and Attitudes towards AI (豹告 | 普通人眼中的AI：大众AI认知、感受、态度调研报告).” Weixin Official Accounts Platform, September 9, 2019. http://mp.weixin.qq.com/s?__biz=MzU3NjI0MzgxOQ==&mid=2247489828&idx=1&sn=2efbbed9c94c8fbb98e5fcbc2700417c&chksm=fd178820ca6001361dc29bacfd925439e06505b3efdcd8475d8a738c21028ac0a5a628a21d2f#rd.

Chen, Xuanting, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. “How Robust Is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks.” arXiv, March 1, 2023. https://doi.org/10.48550/arXiv.2303.00293.

Chengdu Municipal Bureau of Economic and Information Technology (成都市经济和信息化局) and Chengdu New Economic Development Commission (成都市新经济发展委员会). “Notice on the Publication of ‘Several Measures for Accelerating Large Model Innovation and New Applications to Promote the High Quality Development of the AI Industry in Chengdu Municipality’ (关于印发《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》的通知).” Chengdu Municipal Bureau of Economic and Information Technology (成都市经济和信息化局), August 4, 2023. https://cdjx.chengdu.gov.cn/cdsjxw/c160798/2023-08/04/content_ba168b8475ad4419ac6c7393012e2f5c.shtml.

Chern, I.-Chun, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. “FacTool: Factuality Detection in Generative AI -- A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios.” arXiv, July 26, 2023. https://doi.org/10.48550/arXiv.2307.13528.

China Academy of Information and Communications Technology (CAICT) (中国信通院). “CAICT Formally Begins Second Batch of ‘Trustworthy AI’ Evaluations for 2021 (中国信通院2021年第二批'可信AI'评测正式启动)," September 23, 2021. http://www.caict.ac.cn/xwdt/ynxw/202109/t20210923_390249.htm.

China Academy of Information and Communications Technology (CAICT) (中国信通院) and JD Explore Academy (京东探索研究院). "Trustworthy AI White Paper." China Academy of Information and Communications Technology (CAICT) (中国信通院), August 2021. http://www.caict.ac.cn/kxyj/qwfb/bps//202107/P020210709319866413974.pdf.

China Daily. "Testin Cloud Testing Initiates and Participates in the 'New Generation AI Industry Self-Governance Convention' Assisting the Sustainable Development of the AI Industry (Testin云测发起并参与《新一代人工智能行业自律公约》，助力AI行业可持续发展)," August 20, 2019. https://tech.chinadaily.com.cn/a/201908/20/WS5d5b8732a31099ab995da7d8.html.

China Daily. "Why Does Xi Jinping Emphasize 'Bottom-Line Thinking' (习近平为什么强调'底线思维')," January 30, 2019. https://china.chinadaily.com.cn/a/201901/30/WS5c510c53a31010568bdc7697.html.

China Daily. "(Zhongguancun Forum) Baidu CEO Robin Li: Large Models Are about to Change the World (【中关村论坛】百度李彦宏：大模型即将改变世界)," May 26, 2023. https://tech.chinadaily.com.cn/a/202305/26/WS647027a4a3105379893760ab.html.

China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院). "English Introduction (英文介绍)." Accessed October 11, 2023. https://www.cc.cesi.cn/english.aspx.

China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院). "National AI Standardization Overall Group and Expert Consultation Group Created (国家人工智能标准化总体组和专家咨询组成立)," January 19, 2018. http://www.cesi.cn/201801/3539.html.

China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院). "Our Institute Attended the Third Plenary Meeting of ISO/IEC JTC 1/SC 42 (我院参加ISO/IEC JTC 1/SC 42第三次全体会议)," May 6, 2019. http://www.cesi.cn/201905/5057.html.

China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院). "Our Institute Undertook the First Plenary Meeting of the ISO/IEC JTC 1/SC 42 AI Technical Sub-Committee (我院承办ISO/IEC JTC 1/SC 42人工智能分技术委员会第一次全会)," April 23, 2018. http://www.cesi.cn/201804/3821.html.

China Internet Information Center (中国网). "Li Keqiang Discusses AlphaGo: It Does Not Matter Who Wins the Human-Machine Battle, Machines Are Still Created by Humans (李克强谈AlphaGo：人机大战不管输赢 机器还是人造的)," March 16, 2016. http://www.china.com.cn/lianghui/news/2016-03/16/content_38039432.htm.

China Knowledge. "Shanghai 2018 World Artificial Intelligence Conference Kicks off, China's AI Biggest Players Surfaced," September 18, 2018. https://www.chinaknowledge.com/News/DetailNews?id=80279.

China Law Translate. "Interim Measures for the Management of Generative Artificial Intelligence Services." *China Law Translate* (blog), July 13, 2023. https://www.chinalawtranslate.com/generative-ai-interim/.

———. "Provisions on the Administration of Deep Synthesis Internet Information Services." *China Law Translate* (blog), December 12, 2022. https://www.chinalawtranslate.com/deep-synthesis/.

———. "Translation: Provisions on the Management of Algorithmic Recommendations in Internet Information Services." *China Law Translate* (blog), January 4, 2022. https://www.chinalawtranslate.com/algorithms/.

China News (中新网). "Alibaba Group CTO Cheng Li: the science and technology ethics governance committee should be the 'gatekeeper' of technological innovation (阿里集团CTO程立：科技伦理治理委员会要做技术创新的'守

门人’),” September 2, 2022. https://www.sh.chinanews.com.cn/kjjy/2022-09-02/102917.shtml.

Chinese Academy of Sciences (中国科学院). “Zhang Bo: A Founder of Artificial Intelligence in China (张钹：中国人工智能奠基者),” August 30, 2021. https://www.cas.cn/xzfc/202108/t20210830_4803805.shtml.

Chinese Association for Artificial Intelligence (中国人工智能学会). “Introduction to the Chinese Association for Artificial Intelligence (中国人工智能学会简介).” Accessed October 15, 2023. https://www.caai.cn/index.php?s=/home/article/index/id/2.html.

Chris Olah. “Interpretability Dreams.” *Transformer Circuits Thread* (blog), May 24, 2023. https://transformer-circuits.pub/2023/interpretability-dreams/index.html.

Christiano, Paul, Buck Shlegeris, and Dario Amodei. “Supervising Strong Learners by Amplifying Weak Experts.” arXiv, October 19, 2018. https://doi.org/10.48550/arXiv.1810.08575.

CoAI. “CoAI (清华大学交互式人工智能课题组).” Accessed October 12, 2023. https://coai.cs.tsinghua.edu.cn/.

Communist Youth Daily (中国青年报). “77.8% of Respondents Are Optimistic about the Development of Artificial Intelligence in China (77.8%受访者看好我国人工智能的发展),” June 13, 2017. https://zqb.cyol.com/html/2017-06/13/nw.D110000zgqnb_20170613_1-07.htm .

Concordia AI. “AI Safety in China #1.” Substack newsletter. *AI Safety in China* (blog), August 24, 2023. https://aisafetychina.substack.com/p/a6e4bdf0-f687-4ff7-b2c4-dfb06ababd1f.

———. “AI Safety in China #2.” Substack newsletter. *AI Safety in China* (blog), September 7, 2023. https://aisafetychina.substack.com/p/ai-safety-in-china-2.

———. “AI Safety in China #3.” Substack newsletter. *AI Safety in China* (blog), September 20, 2023. https://aisafetychina.substack.com/p/ai-safety-in-china-3.

Concordia AI. “Homepage.” Accessed October 18, 2023. https://concordia-consulting.com/.

Cyberspace Administration of China (网信办). “Global AI Governance Initiative (全球人工智能治理倡议),” October 18, 2023. http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

Cyberspace Administration of China (网信办), Ministry of Industry and Information Technology (工信部), and Ministry of Public Security (公安部). “Provisions on Management of Deep Synthesis in Internet Information Service (互联网信息服务深度合成管理规定),” November 25, 2022. https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm.

Cyberspace Administration of China (网信办), Ministry of Industry and Information Technology (工信部), Ministry of Public Security (公安部), and State Administration of Market Regulation (市场监管总局). “Administrative Provisions on Algorithm Recommendation for Internet Information Services (互联网信息服务算法推荐管理规定).” Cyberspace Administration of China (网信办), January 4, 2022. http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

Cyberspace Administration of China (网信办), National Development and Reform Commission (发改委), Ministry of Education (教育部), Ministry of Science and Technology (科技部), Ministry of Industry and Information Technology (工信部), Ministry of Public Security (公安部), and National Radio and Television Administration (广电总局). “Interim Measures for the Management of Generative Artificial Intelligence Services (生成式人工智能服务管理暂行办法).” Cyberspace Administration of China (网信办), July 13, 2023. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

Cyberspace Administration of China (网信办). “Personal Information Protection Certification Implementation Rules (个人信息保护认证实施规则),” November 18, 2022. http://www.cac.gov.cn/2022-11/18/c_1670399936983876.htm.

Cyberspace Administration of China (网信办). “Regulations for the Security Assessment of Internet Information Services Having Public Opinion

Properties or Social Mobilization Capacity (具有舆论属性或社会动员能力的互联网信息服务安全评估规定)," November 15, 2018. http://www.cac.gov.cn/2018-11/15/c_1123716072.htm.

Dan Harsha. "Long-Term Survey Reveals Chinese Government Satisfaction." *The Harvard Gazette* (blog), July 9, 2020. https://news.harvard.edu/gazette/story/2020/07/long-term-survey-reveals-chinese-government-satisfaction/.

Data Science and AI (数据科学人工智能). "Academician Gao Wen Gave His First Exclusive Interview after Explaining AI Development at the CPC Politburo (高文院士在中共中央政治局讲解人工智能发展后首次接受专访)." Zhihu (知乎), December 2, 2018. https://zhuanlan.zhihu.com/p/51412598.

Deng, Jiawen, Hao Sun, Zhexin Zhang, Jiale Cheng, and Minlie Huang. "Recent Advances towards Safe, Responsible, and Moral Dialogue Systems: A Survey." arXiv, March 6, 2023. http://arxiv.org/abs/2302.09270.

Deng, Jiawen, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. "COLD: A Benchmark for Chinese Offensive Language Detection." arXiv, October 19, 2022. http://arxiv.org/abs/2201.06025.

Department for Science, Innovation and Technology. "UK Government Sets out AI Safety Summit Ambitions." GOV.UK, September 4, 2023. https://www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions.

Department of Computer Science and Technology, Tsinghua University (清华大学计算机科学与技术系). "Jun Zhu's Homepage." Accessed October 19, 2023. https://ml.cs.tsinghua.edu.cn/~jun/index.shtml.

Department of Computer Science and Technology, Tsinghua University (清华大学计算机科学与技术系). "Huang Minlie (黄民烈)," March 31, 2021. https://www.cs.tsinghua.edu.cn/info/1121/5620.htm.

Deyi Xiong (熊德意). "Home." Accessed October 12, 2023. https://dyxiong.github.io/index.html.

Didier Bourguignon. “The Precautionary Principle: Definitions, Applications and Governance.” EPRS | European Parliamentary Research Service, September 12, 2015. https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf.

DigiChina. “Translation: New Rules Target Public Opinion and Mobilization Online in China,” November 21, 2018. https://digichina.stanford.edu/work/new-rules-target-public-opinion-and-mobilization-online-in-china-translation/.

Digital Futures Lab | Konrad-Adenauer-Stiftung. “Reframing AI Governance: Perspectives from Asia,” July 2022. https://assets.website-files.com/62c21546bfcfcd456b59ec8a/62df3bbcd1d3f82534a706f1_%E2%80%A2Report_AI_in_Asia.pdf.

Ding, Jeffrey. “ChinAI #231: Latest SuperCLUE Rankings of Large Language Models.” Substack newsletter. *ChinAI Newsletter* (blog), July 31, 2023. https://chinai.substack.com/p/chinai-231-latest-superclue-rankings.

Dong, Hanze, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. “RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment.” arXiv.org, April 13, 2023. https://arxiv.org/abs/2304.06767v3.

Dong, Yinpeng, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. “How Robust Is Google’s Bard to Adversarial Image Attacks?” arXiv, October 14, 2023. http://arxiv.org/abs/2309.11751.

Dong, Yinpeng, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. “Black-Box Detection of Backdoor Attacks with Limited Information and Data.” arXiv, March 24, 2021. https://doi.org/10.48550/arXiv.2103.13127.

Drexel, Bill, and Hannah Kelley. “China Is Flirting With AI Catastrophe.” *Foreign Affairs*, May 30, 2023. https://www.foreignaffairs.com/china/china-flirting-ai-catastrophe.

European Commission [@EU_Commission]. "Mitigating the Risk of Extinction from AI Should Be a Global Priority. And Europe Should Lead the Way, Building a New Global AI Framework Built on Three Pillars: Guardrails, Governance and Guiding Innovation ↓ Https://T.Co/t7UA9rgN1H." Tweet. *X*, September 14, 2023. https://twitter.com/EU_Commission/status/1702295053668946148.

"Fei Huang." Accessed October 12, 2023. https://sites.google.com/view/fei-huang.

Fu, Yu, Deyi Xiong, and Yue Dong. "Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy." arXiv, July 25, 2023. https://doi.org/10.48550/arXiv.2307.13808.

"Full Text: Proposal of the People's Republic of China on the Reform and Development of Global Governance," September 13, 2023. https://english.news.cn/20230913/edf2514b79a34bf6812a1c372dcdfc1b/c.html?mc_cid=8031f71d00&mc_eid=ccbfb1d564.

Future Forum (未来论坛). "AI Ethics and Governance Series Issue 01 · The Theory and Practice of AI for Good - Future Forum (AI伦理与治理系列01期 · AI向善的理论与实践)," April 28, 2021. http://www.futureforum.org.cn/cn/nav/detail/528.html.

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

Gabriel, Iason. "Artificial Intelligence, Values and Alignment." *Minds and Machines* 30, no. 3 (September 2020): 411–37. https://doi.org/10.1007/s11023-020-09539-2.

GAIR. "GAIR: Generative Artificial Intelligence Research Lab." Accessed October 12, 2023. https://plms.ai/.

"GAIR-NLP/Factool." Python. 2023. Reprint, Generative Artificial Intelligence Research Lab (GAIR), October 10, 2023. https://github.com/GAIR-NLP/factool.

General Office of Beijing Municipal People's Government (北京市人民政府办公厅). "Notice by the General Office of the Beijing Municipal People's Government on the Publication of the ‘Several Measures for Promoting the Innovation and Development of Artificial General Intelligence in Beijing’ (北京市人民政府办公厅关于印发《北京市促进通用人工智能创新发展的若干措施》的通知),” May 30, 2023. https://www.beijing.gov.cn/zhengce/gfxwj/202305/t20230530_3116869.html.

“General Office of the Communist Party of China and General Office of the State Council on Publishing ‘Opinions On Strengthening Science and Technology Ethics Governance’ (中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》),” March 20, 2022. https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

“Global Opinions and Expectations About Artificial Intelligence: A Global Advisor Survey.” Ipsos, January 2022. https://www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Global-opinions-and-expectations-about-AI-2022.pdf.

Google Scholar. “Yiming Li (李一鸣).” Accessed October 19 2023. https://scholar.google.com/citations?hl=zh-CN&user=mSW7kU8AAAAJ&view_op=list_works.

Graham Webster. “Translation: Chinese AI Alliance Drafts Self-Discipline ‘Joint Pledge.’” New America, June 17, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/.

Graham Webster and Lorand Laskai. “Translation: Chinese Expert Group Offers ‘Governance Principles’ for ‘Responsible AI.’” *DigiChina* (blog), June 17, 2019. https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/.

Graham Webster, Rogier Creemers, Elsa Kania, and Paul Triolo. “Full Translation: China’s ‘New Generation Artificial Intelligence Development Plan’ (2017).” *DigiChina* (blog), August 1, 2017. https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

Gui, Tao, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, et al. “TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing.” arXiv, May 5, 2021. http://arxiv.org/abs/2103.11441.

Guozha Zhang (张国祚). “Discussing ‘Bottom-Line Thinking’ (谈谈‘底线思维’),” October 1, 2013. https://news.12371.cn/2013/10/01/ARTI1380592471362492.shtml?from=groupmessage&isappinstalled=0.

He, Conghui, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. “WanJuan: A Comprehensive Multimodal Dataset for Advancing English and Chinese Large Models.” arXiv, September 15, 2023. http://arxiv.org/abs/2308.10755.

Hern, Alex, and Kiran Stacey. “No 10 Acknowledges ‘Existential’ Risk of AI for First Time.” *The Guardian*, May 25, 2023, sec. Technology. https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak.

Hongqiao Lyu (吕红桥). “[Frontier] How to View AI Threat Theory (【前沿】如何看待‘人工智能威胁论’？).” Weixin Official Accounts Platform, April 27, 2017. http://mp.weixin.qq.com/s?__biz=MzA4MjA1NjAzMQ==&mid=2651099910&idx=2&sn=34f9b4b0a89702b6945e66b492a18f79&chksm=847bb536b30c3c2048304dc64b7be0d1f9ba0cd403d97987eada34b330eb6c0a89a59fc12286#rd.

Huaihong He (何怀宏). *Does Humanity Still Have A Future? (人类还有未来吗).* Guangxi Normal University Press (广西师范大学出版社), 2020. https://book.douban.com/subject/35197706/.

———. “Huaihong HE.” Translated by Concordia AI. Chinese Perspectives on AI. Accessed October 12, 2023. https://chineseperspectives.ai/Huaihong-HE.

Huang, Yufei, and Deyi Xiong. “CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models.” arXiv, June 28, 2023. https://doi.org/10.48550/arXiv.2306.16244.

Huawei. “AI Security (AI安全).” Accessed October 13, 2023. https://www.huawei.com/cn/trust-center/ai-section.

Huawei. “AI Security White Paper (AI安全白皮书).” Accessed October 19, 2023. https://www-file.huawei.com/-/media/corporate/pdf/cyber-security/ai-security-white-paper-cn.pdf.

Huawei Noah’s Ark Lab. “About.” Accessed October 12, 2023. https://noahlab.com.hk/#/about.

Huawei Noah’s Ark Lab. “Research.” Accessed October 12, 2023. https://noahlab.com.hk/#/research.

Hui Shen (沈慧). “Tan Tieniu: The Overall Development Level of Artificial Intelligence Is Still in an Early Stage (谭铁牛：人工智能总体发展水平仍然处于起步阶段).” Baidu, July 26, 2018. https://baijiahao.baidu.com/s?id=1607015221234758722&wfr=spider&for=pc.

ICloudnews. “2019 World AI Conference Safety/Security High Level Discussion Is about to Start (2019世界人工智能安全高端对话即将开启),” August 23, 2019. https://www.icloudnews.net/a/22912.html.

Industrial Department II of the Standardization Administration of China (国家标准化管理委员会工业二部) and China Electronics Standardization Institute. “Artificial Intelligence Standardization White Paper (2018 Edition) (人工智能标准化白皮书).” China Electronic Standardization Institute (CESI) (中国电子技术标准化研究院), January 2018. http://www.cesi.cn/images/editor/20180124/20180124135528742.pdf.

Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院). “Highlights | 2022 Beijing Academy of AI Conference-AI Ethics, Governance, and Sustainable Development Forum Successfully Held (精彩观点一览 ｜ 2022北京智源大会-人工智能伦理、治理与可持续发展分论坛成功召开),” June 7, 2022. https://aiig.tsinghua.edu.cn/info/1294/1510.htm.

Institute of Automation, Chinese Academy of Sciences (中国科学院自动化研究院). “China Artificial Industry Industry Alliance (AIIA) Founded, Institute of Autonomation Selected as a Vice-Chairman Institution (中国人工智能产业

发展联盟成立，自动化所当选副理事长单位),” October 13, 2017. http://www.ia.cas.cn/xwzx/ttxw/201710/t20171013_4873206.html.

International Research Center for AI Ethics and Governance. “Beijing Artificial Intelligence Principles.” Accessed October 13, 2023. https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/.

InternLM Team. “InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities.” GitHub, June 3, 2023. https://github.com/InternLM/InternLM-techreport/blob/main/InternLM.pdf.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. “AI Safety via Debate.” arXiv, October 22, 2018. https://doi.org/10.48550/arXiv.1805.00899.

ISO. “ISO/IEC JTC 1/SC 42 - Artificial Intelligence.” Accessed October 11, 2023. https://www.iso.org/committee/6794475.html.

Jamie Elsey and David Moss. “US Public Opinion of AI Policy and Risk.” Rethink Priorities, May 12, 2023. https://rethinkpriorities.org/publications/us-public-opinion-of-ai-policy-and-risk.

Jason Si. “These Rules Could Save Humanity from the Threat of Rogue AI.” World Economic Forum, May 8, 2019. https://www.weforum.org/agenda/2019/05/these-rules-could-save-humanity-from-the-threat-of-rogue-ai/.

Jay Solomon. “Ex-Pentagon Chief: US, China Locked in Existential Struggle for AI Dominance | Semafor.” Semafor, July 28, 2023. https://www.semafor.com/article/07/27/2023/us-china-existential-struggle-ai-dominance.

Jeffrey Ding and Jenny Xiao. “Recent Trends in China’s Large Language Model Landscape.” Centre for the Governance of AI, April 28, 2023. https://cdn.governance.ai/Trends_in_Chinas_LLMs.pdf.

Ji, Jiaming, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. "OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research." arXiv, May 16, 2023. http://arxiv.org/abs/2305.09304.

Jia Liu (刘佳). "China Already Has 79 Large Models with 1 Billion Parameters, and Industry Is Calling for Establishing a 'Moat' for Independent Innovation as Quickly as Possible (中国已有79个10亿参数大模型，业界呼吁尽快建立自主创新'护城河')." Yicai (第一财经), May 29, 2023. https://m.yicai.com/news/101769137.html.

Jie Fu (付杰). "Big AI Dream | Jie Fu." Big AI Dream | Jie Fu. Accessed October 12, 2023. https://bigaidream.github.io/.

Jinxu Wang (王金许). "Huang Tiejun: 'Intelligence for Use, Machine for the Body,' Realizing Artificial Brains within 30 Years (黄铁军：'智能为用，机器为体'，30年内实现人造大脑)." Leiphone (雷锋网), January 12, 2018. https://www.leiphone.com/category/ai/o67S8c36efqqZW32.html.

Jun Zhang (张军). "Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security." Permanent Mission of the People's Republic of China to the UN (中华人民共和国常驻联合国代表团), July 18, 2023. http://un.china-mission.gov.cn/eng/hyyfy/202307/t20230719_11114947.htm.

KEG. "Jie Tang (Tang, Jie) 唐杰," Accessed October 19, 2023. https://keg.cs.tsinghua.edu.cn/jietang/.

Knowledge Engineering Group (KEG) & Data Mining at Tsinghua University. "ChatGLM-6B/README.Md at Main · THUDM/ChatGLM-6B." GitHub, April 25, 2023. https://github.com/THUDM/ChatGLM-6B/blob/main/README.md.

Kwan Yee Ng, Jason Zhou, Ben Murphy, Rogier Creemers, and Hunter Dorwart. "Translation: Artificial Intelligence Law, Model Law v. 1.0 (Expert Suggestion Draft) – Aug. 2023." *DigiChina* (blog), August 23, 2023. https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/.

Lao Wang (老王). “China’s Most Secret Research Base-Huawei 2012 Lab (中国最神秘的研究基地——华为2012实验室),” August 10, 2016. https://www.leiphone.com/category/industrynews/5fWci6bJoL7JW5Wr.html.

Lee Kai-Fu (李開復) and Yonggang Wang (王詠剛). *AI Is Here (人工智慧來了)*. 天下文化, 2017. https://web.archive.org/web/20231019060345/https://www.books.com.tw/products/0010750425.

Li Keqiang. “Report on the Work of the Government.” The State Council, March 16, 2017. https://english.www.gov.cn/premier/news/2017/03/16/content_281475597911192.htm.

Li, Linyang, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. “BERT-ATTACK: Adversarial Attack Against BERT Using BERT.” arXiv, October 1, 2020. http://arxiv.org/abs/2004.09984.

Li, Quanyi, Zhenghao Peng, Haibin Wu, Lan Feng, and Bolei Zhou. “Human-AI Shared Control via Policy Dissection.” In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. https://openreview.net/forum?id=LCOv-GVVDkp.

Li, Yiming. “Backdoor Learning Resources.” Accessed October 18, 2023. https://github.com/THUYimingLi/backdoor-learning-resources.

Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. “Backdoor Learning: A Survey.” arXiv, February 16, 2022. http://arxiv.org/abs/2007.08745.

Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. “Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models’ Alignment.” arXiv, August 10, 2023. http://arxiv.org/abs/2308.05374.

Luiza Ch. Savage and Nancy Scola. “‘We Are Being Outspent. We Are Being Outpaced’: Is America Ceding the Future of AI to China?” POLITICO, July 18, 2019.

https://www.politico.com/story/2019/07/18/global-translations-ai-china-1598442.

Ma, Yi, Doris Tsao, and Heung-Yeung Shum. "On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence." arXiv, July 27, 2022. https://doi.org/10.48550/arXiv.2207.04630.

Matt Schiavenza. "China's 'Sputnik Moment' and the Sino-American Battle for AI Supremacy." Asia Society, September 25, 2018. https://asiasociety.org/blog/asia/chinas-sputnik-moment-and-sino-american-battle-ai-supremacy.

Matt Sheehan and Jacob Feldgoise. "What Washington Gets Wrong About China and Technical Standards." Carnegie Endowment for International Peace, February 27, 2023. https://carnegieendowment.org/2023/02/27/what-washington-gets-wrong-about-china-and-technical-standards-pub-89110.

Matt Sheehan and Sharon Du. "How Food Delivery Workers Shaped Chinese Algorithm Regulations." Carnegie Endowment for International Peace, November 2, 2022. https://carnegieendowment.org/2022/11/02/how-food-delivery-workers-shaped-chinese-algorithm-regulations-pub-88310.

———. "What China's Algorithm Registry Reveals about AI Governance." Carnegie Endowment for International Peace, December 9, 2022. https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606.

Mazzocco, Ilaria, and Scott Kennedy. "Public Opinion in China: A Liberal Silent Majority?," February 9, 2022. https://www.csis.org/analysis/public-opinion-china-liberal-silent-majority.

McFadden, Mark, Kate Jones, Emily Taylor, and Georgia Osborn. "Harmonising Artificial Intelligence: The Role of Standards in the EU AI Regulation." Oxford Information Labs, December 2021.

https://oxil.uk/publications/2021-12-02-oxford-internet-institute-oxil-harmonising-ai/Harmonising-AI-OXIL.pdf.

"Megvii's 'AI Application Guidelines' Full Document Is Published, Advocating for Beneficial AI Technology (旷视《人工智能应用准则》全文公布 提倡善用AI技术)." 36kr, July 17, 2019. https://36kr.com/p/1724031811585.

Microsoft Research. "Responsible AI Workshop: An Interdisciplinary Approach," October 24, 2022. https://www.microsoft.com/en-us/research/event/responsible-ai-an-interdisciplinary-approach-workshop/.

Microsoft Research. "Xing Xie at Microsoft Research." Accessed October 12, 2023. https://www.microsoft.com/en-us/research/people/xingx/.

Ministry of Foreign Affairs (外交部). "Foreign Ministry Spokesperson's Remarks on the Global AI Governance Initiative," October 18, 2023. https://www.fmprc.gov.cn/eng/xwfw_665399/s2510_665401/202310/t20231018_11162874.html.

Ministry of Foreign Affairs (外交部). "The Global Security Initiative Concept Paper," February 21, 2023. https://www.fmprc.gov.cn/mfa_eng/wjbxw/202302/t20230221_11028348.html.

Ministry of Foreign Affairs (外交部). "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)," December 14, 2021. https://www.fmprc.gov.cn/eng/wjdt_665385/wjzcs/202112/t20211214_10469512.html.

Ministry of Foreign Affairs (外交部). "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)," November 17, 2022. https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html.

Ministry of Science and Technology (科技部). "'Ethical Norms for New Generation AI' Published (《新一代人工智能伦理规范》发布)," September 26, 2021. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

Ministry of Science and Technology (科技部), Ministry of Education (教育部), Ministry of Industry and Information Technology (工信部), Ministry of Agriculture and Rural Affairs (农业部), National Health Commission (国家卫生健康委), Chinese Academy of Sciences (中国科学院), Chinese Academy of Social Sciences (中国社科院), Chinese Academy of Engineering (中国工程院), China Association for Science and Technology (中国科协), and Central Military Commission Science and Technology Committee (中央军委科技委). "Notice on the Publishing of the 'Science and Technology Ethics Review Plan (Trial)' (关于印发《科技伦理审查办法（试行）》的通知)," October 8, 2023. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html.

MSRA. "Xie Xing: conduct research that can withstand the test of time, develop responsible AI (谢幸：做经得起时间检验的研究，打造负责任的人工智能)," August 9, 2023. https://www.msra.cn/zh-cn/news/people-stories/xing-xie-societal-ai.

Nanfang Daily (南方日报). "Guangdong enterprises take the lead in initiating an AI industry self-discipline convention, covering 8 aspects including privacy protection (粤企领衔发起AI行业自律公约，涉及隐私保护等8个方面)," August 19, 2019. https://xapp.southcn.com/node_fb07388412?url=https://law.southcn.com/node_d384a70bd7/7e7932524c.shtml&is_app=1.

National AI Standardization Overall Group (国家人工智能标准化总体组) and National Information Technology Standardization Committee AI Subcommittee (全国信标委人工智能分委会). "AI Ethics Governance Standardization Guide, 2023 Version (人工智能伦理治理标准化指南, 2023版)," March 2023. https://web.archive.org/web/20230531193844/https://www.aipubservice.com/airesource/fs/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BC%A

6%E7%90%86%E6%B2%BB%E7%90%86%E6%A0%87%E5%87%86%E5%8C%96%E6%8C%87%E5%8D%97.pdf.

National Information Security Standardization Technical Committee's (TC260) Big Data Security Special Working Group (全国信息安全标准化技术委员会大数据安全标准特别工作组). "AI Safety/Security Standardization White Paper - 2019 Version (人工智能安全标准化白皮书 - 2019 版)," October 2019. http://www.cesi.cn/images/editor/20191101/20191101115151443.pdf.

———. "AI Safety/Security Standardization White Paper - 2023 Version (人工智能安全标准化白皮书 - 2023 版)," May 2023. https://www.tc260.org.cn/front/postDetail.html?id=20230531105159.

National Natural Science Foundation of China (国家自然科学基金委员会). "Announcement on the Publication of 2023 Project Guidelines for the Important Research Plan on Interpretable and Generalizable New Generation AI Methods (关于发布可解释、可通用的下一代人工智能方法重大研究计划2023年度项目指南的通告)," April 3, 2023. https://www.nsfc.gov.cn/publish/portal0/tab434/info89087.htm.

National New Generation AI Governance Expert Committee (国家新一代人工智能治理专业委员会). "Develop Responsible AI: New Generation AI Governance Principles Published (发展负责任的人工智能：新一代人工智能治理原则发布)." Ministry of Science and Technology (科技部), June 17, 2019. https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html.

Neural Reality (神经现实). "The Theory and Practice of AI for Good (AI向善的理论与实践)." NetEase (网易), October 16, 2023. https://www.163.com/dy/article/GBTR244R0512M9G9.html.

New Governance (新治理). "'AI Law (Model Law) 1.0' (Expert Suggestion Draft) Drafting Statement and Full Document (《人工智能法（示范法）1.0》（专家建议稿）起草说明和全文)." Weixin Official Accounts Platform, August 15, 2023. http://mp.weixin.qq.com/s?__biz=Mzg2NDYzMzMyMA==&mid=2247485974&idx=1&sn=f543b06dc59b3e81dfb0b45ad744b6d1&chksm=ce672291f910ab87083938537f71a4d7b4e313d00365d8a8a69c493b11233388ea23d6c1cb7b#rd.

Nicolas Boyon. “Opinions about AI Vary Depending on Countries’ Level of Economic Development.” Ipsos, January 5, 2022. https://www.ipsos.com/en/global-opinions-about-ai-january-2022.

NIST. “AI Risk Management Framework.” Accessed October 11, 2023. https://www.nist.gov/itl/ai-risk-management-framework.

OECD. *Understanding and Applying the Precautionary Principle in the Energy Transition*. Paris: Organisation for Economic Co-operation and Development, 2023. https://www.oecd-ilibrary.org/governance/understanding-and-applying-the-precautionary-principle-in-the-energy-transition_5b14362c-en.

OECD.ai Policy Observatory. “G20 AI Principles,” June 9, 2019. https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf.

Ó hÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. “Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance.” *Philosophy & Technology* 33, no. 4 (December 1, 2020): 571–93. https://doi.org/10.1007/s13347-020-00402-x.

OpenAI. “Frontier Model Forum.” OpenAI, July 26, 2023. https://openai.com/blog/frontier-model-forum.

———. “GPT-4 System Card,” March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

PAIR Lab: PKU Alignment and Interaction Research Lab. “PAIR Lab: PKU Alignment and Interaction Research Lab.” Accessed October 12, 2023. https://pair-lab.com/.

Pan, Alexander, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. “Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark.” arXiv, June 12, 2023. https://doi.org/10.48550/arXiv.2304.03279.

Pan, Yunhe. “Heading toward Artificial Intelligence 2.0.” *Engineering* 2, no. 4 (December 1, 2016): 409–13. https://doi.org/10.1016/J.ENG.2016.04.018.

The Paper. “Why was AlphaGo born in Silicon Valley instead of China? He told the real reason” (“为什么AlphaGo生在硅谷，而非中国？他说出了真实原因)” August 31, 2019. https://m.thepaper.cn/wifiKey_detail.jsp?contid=1779773&from=wifiKey#.

Pedro A. Ortega, Vishal Maini, and the DeepMind safety team. “Building Safe Artificial Intelligence: Specification, Robustness, and Assurance.” *Medium* (blog), September 27, 2018. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.

Pengfei Liu. “Pengfei Liu.” Accessed October 12, 2023. http://pfliu.com/.

People.cn (人民网). “Ministry of Foreign Affairs Publishes ‘Proposal of the People’s Republic of China on the Reform and Development of Global Governance’ (外交部发布《关于全球治理变革和建设的中国方案》),” September 13, 2023. http://world.people.com.cn/n1/2023/0913/c1002-40077078.html.

PKU-Alignment. “PKU Beaver: Constrained Value-Aligned LLM via Safe RLHF.” Accessed October 10, 2023. https://pku-beaver.github.io/.

———. “BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset.” Accessed October 10, 2023. https://sites.google.com/view/pku-beavertails/home.

Politburo of the Communist Party of China (中共中央政治局). “Analyzing and Researching the Present Economic Situation and Economic Work (分析研究当前经济形势和经济工作).” Chinese Government, April 28, 2023. https://www.gov.cn/yaowen/2023-04/28/content_5753652.htm.

“Pre-Research Report: Risks and Governance of AI from the Perspective of Chinese Youth.” Center for International Security and Strategy Tsinghua University (清华大学战略与安全研究中心), 2019. https://ciss.tsinghua.edu.cn/upload_files/atta/1589025522890_83.pdf.

QbitAI (量子位). "Turing Award Winner, 200+ Leading AI Academics, 30+ Special Forums, the Annual Grand Event Is Here (图灵奖得主、200+AI顶尖学术领袖，30+专题论坛，年度盛会来了)." Matpool, May 18, 2021. https://matpool.com/blog/60a5c2b2c5695302acca422c/.

Qiao-Franco, Guangyu, and Rongsheng Zhu. "China's Artificial Intelligence Ethics: Policy Development in an Emergent Community of Practice." *Journal of Contemporary China* 0, no. 0 (2022): 1–17. https://doi.org/10.1080/10670564.2022.2153016.

Qiqi Gao (高奇琦). "Artificial Intelligence, the Fourth Industrial Revolution, and the International Political-Economic Landscape (人工智能, 四次工业革命与国际政治经济格局)." 当代世界与社会主义, no. 6 (2019): 12–19.

Quanshi Zhang (张拳石). "Curriculum Vitae Quanshi Zhang," Accessed October 12, 2023. http://qszhang.com/files/CV.pdf.

———. "Publications | Quanshi Zhang." Accessed October 12, 2023. http://qszhang.com/index.php/publications/.

Qun Liu (刘群). "LIU, Qun (刘群)." Accessed October 12, 2023. https://liuquncn.github.io/index_en.html.

Quora. "Is AI an Existential Threat to Humanity?" Accessed October 12, 2023. https://www.quora.com/Is-AI-an-existential-threat-to-humanity.

Qwen Team, Alibaba Group. "Qwen Technical Report," Accessed October 12, 2023. https://qianwen-res.oss-cn-beijing.aliyuncs.com/QWEN_TECHNICAL_REPORT.pdf.

RealAI. "About Us: Mission and Vision (关于我们：使命愿景)." Accessed October 18, 2023. https://www.realai.ai/about?scrollId=t19.

RealAI. "AI Security Platform: RealSafe (人工智能安全平台RealSafe)." Accessed October 19, 2023. https://www.realai.ai/products/55.html.

Renhan Li (李仁涵) and Qingqiao Huang (黄庆桥). *AI and Values (人工智能与价值观)*. Shanghai Jiaotong University Press (上海交通大学出版社), 2021. https://weread.qq.com/web/bookDetail/cdd320107260a44acdd2189.

Reuters. "China Has Won AI Battle with U.S., Pentagon's Ex-Software Chief Says." October 11, 2021, sec. Technology. https://www.reuters.com/technology/united-states-has-lost-ai-battle-china-pentagons-ex-software-chief-says-2021-10-11/.

Reza Hasmath. "How China Sees the World in 2023." The China Institute, University of Alberta, May 2023. https://www.ualberta.ca/china-institute/media-library/media-gallery/research/research-papers/2023-china-survey-report/howchinaseestheworld2023.pdf.

Robin Li, "Foundation Models Are Changing the World | Robin Li at the 2023 ZGC Forum." May 2023. https://www.youtube.com/watch?v=-ASsYLzsSxs.

Rogier Creemers and Graham Webster. "Translation: Internet Information Service Deep Synthesis Management Provisions (Draft for Comment) – Jan. 2022." *DigiChina* (blog), February 4, 2022. https://digichina.stanford.edu/work/translation-internet-information-service-deep-synthesis-management-provisions-draft-for-comment-jan-2022/.

———. "Translation: Personal Information Protection Law of the People's Republic of China - Effective Nov. 1, 2021." *DigiChina* (blog), August 20, 2021. https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/.

Rogier Creemers, Graham Webster, and Helen Toner. "Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022." *DigiChina* (blog), January 10, 2022. https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/.

Rogier Creemers, Johanna Costigan, Paul Triolo, Tom Nunlist, Lauren Dudley, Mei Danowski, Martin Chorzempa, Karman Lucero, and Seaton Huang. "Is China's Tech 'Crackdown' or 'Rectification' Over?" *DigiChina* (blog), January 25,

2023. https://digichina.stanford.edu/work/is-chinas-tech-crackdown-or-rectification-over/.

Roose, Kevin. “A.I. Poses ‘Risk of Extinction,’ Industry Leaders Warn.” *The New York Times*, May 30, 2023, sec. Technology. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

Sabine Neschke, Jeremy Pesner, and John Soroushian. “Artificial Intelligence Policy and the European Union: A Look Across the Atlantic.” Bipartisan Policy Center, August 2022. https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2022/08/AI-Policy-and-EU-Final-Paper.pdf.

SenseTime (商汤). “‘Balanced Development’ AI Governance White Paper (‘平衡发展’的人工智能治理白皮书),” September 2022. https://oss.sensetime.com/20221019/d6fb83bdf9e069c70890d7d85214b23e/%E5%B9%B3%E8%A1%A1%E5%8F%91%E5%B1%95%E7%9A%84AI%E6%B2%BB%E7%90%86%E7%99%BD%E7%9A%AE%E4%B9%A6.pdf.

SenseTime (商汤). “SenseTime AI Ethics and Governance Committee Introduction (商汤科技人工智能伦理与治理委员会简介),” October 14, 2022. https://www.sensetime.com/cn/ethics-detail/56847?categoryId=32429.

SenseTime (商汤). “SenseTime Feng Tian: AIGC and Emerging Large Model Risk Governance Technology (商汤田丰：AIGC与大模型风险治理技术涌现),” March 6, 2023. https://www.sensetime.com/cn/ethics-detail/51166748?categoryId=32738.

SenseTime (商汤). “SenseTime S&T AI Ethics and Governance Frontier Practices (商汤科技人工智能伦理与治理前沿实践),” 2022. https://oss.sensetime.com/20221123/e073a90ac65fb9dadc03b7e33a94d414/%E5%95%86%E6%B1%A4%E7%A7%91%E6%8A%80AI%E4%BC%A6%E7%90%86%E4%B8%8E%E6%B2%BB%E7%90%86%E5%89%8D%E6%B2%BF%E5%AE%9E%E8%B7%B5-PDF%E6%8A%A5%E5%91%8A.pdf.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “About Us (关于我们).” Accessed October 12, 2023. https://www.shlab.org.cn/aboutus.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “JR-Large Model Trustworthy AI Algorithms Intern Researcher (JR-大模型可信AI算法见习研究员),” October 7, 2023. https://www.shlab.org.cn/joinus/detail/f1b9cebf-3942-4a66-acd3-1613eb28173f.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Large Model Safety Evaluations-Large Model Safety Youth Researcher (大模型安全评测-大模型安全青年研究员),” August 2, 2023. https://www.shlab.org.cn/joinus/detail/16309e3d-ee57-43d7-88d5-f52e02f3247c?mode=social.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Large Model Safety/Security Evaluations-Large Model Safety/Security Engineer (大模型安全评测-大模型安全工程师),” September 14, 2023. https://www.shlab.org.cn/joinus/detail/653da0a1-9f39-4c74-9b1f-4e6b89053356?mode=campus&keyword=&zhinengId=&commitment=.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Pushi Open Source System - Large Language Model Young Researcher (浦视开源体系-大语言模型青年研究员),” October 11, 2023. https://www.shlab.org.cn/joinus/detail/44480b3d-fbc7-4a41-8ddb-592d46e12f82?mode=social.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Research (研究方向).” Accessed October 12, 2023. https://www.shlab.org.cn/research.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Shanghai AI Lab Pushi Open-Source System Team | Global Recruiting (上海人工智能实验室浦视开源体系团队 | 全球招聘),” August 17, 2023. https://www.shlab.org.cn/news/5443474.

Shanghai Artificial Intelligence Laboratory (上海人工智能实验室). “Shanghai AI Lab Selected as the Leader of the National AI Standardization Overall Group’s

Large Model Special Group 上海人工智能实验室当选国家人工智能标准化总体组大模型专题组组长." Accessed October 11, 2023. https://www.shlab.org.cn/news/5443434.

Shanghai Municipal People's Government (上海市人民政府). "Measures for the Promotion of AI Industry Development in Shanghai Municipality (上海市促进人工智能产业发展条例)," October 1, 2022. https://www.shanghai.gov.cn/hqcyfz2/20230627/3a1fcfeff9234e8e9e6623eb12b49522.html.

Shen, Tianhao, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. "Large Language Model Alignment: A Survey." arXiv, September 26, 2023. http://arxiv.org/abs/2309.15025.

Shen, Wen, Lei Cheng, Yuxiao Yang, Mingjie Li, and Quanshi Zhang. "Can the Inference Logic of Large Language Models Be Disentangled into Symbolic Concepts?" arXiv, April 3, 2023. https://doi.org/10.48550/arXiv.2304.01083.

Shenzhen Municipal People's Congress (深圳市人大常委会). "Shenzhen Special Economic Zone AI Industry Promotion Measures (深圳经济特区人工智能产业促进条例)," September 9, 2022. http://www.szrd.gov.cn/szrd_zlda/szrd_zlda_flfg/flfg_szfg/content/post_834707.html.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, et al. "Model Evaluation for Extreme Risks." arXiv, September 22, 2023. https://doi.org/10.48550/arXiv.2305.15324.

Sina. "Robin Li: Adhere to Four Principles for AI Ethics, Create Responsible, Sustainable AI (李彦宏：坚守人工智能伦理四原则 打造负责、可持续AI)," December 27, 2021. https://finance.sina.com.cn/tech/2021-12-27/doc-ikyakumx6675750.shtml.

SJTU interpretable ML Lab. "Lab for Interpretability and Theory-Driven Deep Learning." Accessed October 15, 2023. https://sjtu-xai-lab.github.io/.

Sogou Encyclopedia (搜狗百科). "Bottom-Line Linking (底线思维)." Accessed October 16, 2023.

https://baike.sogou.com/v59638837.htm?fromTitle=%E5%BA%95%E7%BA%BF%E6%80%9D%E7%BB%B4%EF%BC%88%E4%B8%93%E4%B8%9A%E6%9C%AF%E8%AF%AD%EF%BC%89.

Standardization Administration of China (国家标准委), Cyberspace Administration of China (网信办), National Development and Reform Commission (发改委), Ministry of Science and Technology (科技部), and Ministry of Industry and Information Technology (工信部). "Notice by Five Departments on the Publication of the 'Standardization Construction Guide for National New Generation AI' (五部门关于印发《国家新一代人工智能标准体系建设指南》的通知)," July 27, 2020. https://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm.

Stanford University Human-Centered Artificial Intelligence (HAI). "Global AI Vibrancy Tool: Who's Leading the Global AI Race?" Accessed October 11, 2023. https://aiindex.stanford.edu/vibrancy/.

State Administration of Market Regulation (市场监管总局) and Standardization Administration of China (国家标准委). "Information Security Technology-Assessment Specification for Security of Machine Learning Algorithms (GB/T 42888-2023) (信息安全技术-机器学习算法安全评估规范)," August 6, 2023. http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=E7170BA58AE37AACF4170242EFD25183.

State Council (国务院). "Announcement by the State Council General Office on Publishing the 2023 State Council Yearly Legislative Plan (国务院办公厅关于印发国务院2023年度立法工作计划的通知)," June 6, 2023. https://www.gov.cn/zhengce/content/202306/content_6884925.htm.

State Council (国务院). "The 'Beijing AI Principles' Holds up 'Harmonious and Optimized Coexistence' (人工智能'北京共识'提出'和谐与优化共生')," May 26, 2019. https://www.gov.cn/xinwen/2019-05/26/content_5394829.htm.

State Council (国务院). "Notice by the State Council on the Publication of the New Generation AI Development Plan (国务院关于印发新一代人工智能发展规

划的通知),” July 20, 2017. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

Sun, Hao, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. “On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark.” arXiv, April 4, 2022. https://doi.org/10.48550/arXiv.2110.08466.

Sun, Hao, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. “Safety Assessment of Chinese Large Language Models.” arXiv, April 20, 2023. https://doi.org/10.48550/arXiv.2304.10436.

Sun, Hao, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. “MoralDial: A Framework to Train and Evaluate Moral Dialogue Systems via Moral Discussions.” arXiv, May 26, 2023. http://arxiv.org/abs/2212.10720.

Synced Review (机器之心). “Alibaba Announces the Establishment of Artificial Intelligence Governance and Sustainable Development Laboratory (阿里巴巴宣布成立人工智能治理与可持续发展实验室),” July 13, 2021. https://www.jiqizhixin.com/articles/2021-07-13-13.

Synced Review (机器之心). “(Shanghai Jiaotong University Associate Professor Liu Pengfei is recruiting NLP and Generative AI undergraduate/graduate students for his lab (上海交大副教授刘鹏飞实验室招收NLP、生成式AI方向本科生/研究生),” May 10, 2023. http://posts.careerengine.us/p/645c0068cc3d344e6a1bcc62.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf Doubleday Publishing Group, 2018.

The China Project. “China’s Big Tech Crackdown: A Complete Timeline.” Accessed October 11, 2023. https://thechinaproject.com/big-tech-crackdown-timeline/.

The Fudan Lab For Natural Language Processing Group. “The Fudan Lab For Natural Language Processing Group.” Accessed October 12, 2023. https://nlp.fudan.edu.cn/nlpen/main.htm.

The Institute for the History of Natural Sciences, Chinese Academy of Sciences (中国科学院自然科学史研究所). “Liu Yidong (刘益东).” Accessed October 12, 2023. http://sourcedb.ihns.cas.cn/cn/ihnsexport/200906/t20090602_253811.html.

The National Museum of Nuclear Science & History. “Russell-Einstein Manifesto.” Accessed October 13, 2023. https://ahf.nuclearmuseum.org/ahf/key-documents/russell-einstein-manifesto/.

The Paper (澎湃). “20 Measures! Chengdu Accelerates the Innovation and Application of Large Models to Promote the High Quality Development of the AI Industry (20条措施！成都加快大模型创新应用推进人工智能产业高质量发展),” August 7, 2023. https://m.thepaper.cn/newsDetail_forward_24145627.

Tingyang Zhao (赵汀阳). “Translation: Zhao Tingyang: ‘Near-Term Worries’ and ‘Long-Term Concerns’ of the Artificial Intelligence ‘Revolution’: An Analysis of Ethics and Ontology.” Translated by Jeffrey Ding. Google Docs. Accessed October 12, 2023. https://docs.google.com/document/d/1b9n1IKvMF6kj1NTwd-mP-IvOGFbe3bzqQPOPD-7BVRE/edit.

TJUNLP. “TJUNLP Lab.” Accessed October 12, 2023. https://tjunlp-lab.github.io/.

Tong Zhang (张潼). “Artificial Intelligence Will Eventually Surpass Human Experts (人工智能终将超越人类专家).” Tencent Research Institute (腾讯研究院), August 18, 2017. https://www.tisi.org/15917.

Toutiao (今日头条). “AI Threat Theory: Stubborn Thinking, Doomed Joke (人工智能威胁论：僵化的思考，注定的笑柄),” July 7, 2017. https://www.toutiao.com/article/6439832401802691073/?wid=1697164792407.

Trustworthy AI Evaluations (可信AI评测). “CAICT’s 8th Batch of ‘Trustworthy AI’ Evaluations in 2023 Formally Begins (中国信通院2023年‘可信AI’（第八批）评测正式启动).” Weixin Official Accounts Platform, February 17, 2023. http://mp.weixin.qq.com/s?__biz=Mzg3ODU5NDI0MQ==&mid=2247487529

&idx=2&sn=eeef8e2f145ccb8ee8e248bec93725b8&chksm=cf100187f86788911a4c0d1264349b0e431d5edb1ef17b56bbbeb4a96935dcae0d1318ca9c8b#rd.

Tsinghua Institute for AI International Governance (清华大学人工智能国际治理研究院). "A Thousand People Jointly Issued a 'Soul Questioning': AI Is Raging, Should Humans 'Step on the Brakes' or 'Step on the Accelerator'? (千人联名发出'灵魂拷问'：AI狂飙，人类应该'踩刹车'还是'踩油门'？)." Weixin Official Accounts Platform, March 31, 2023. http://mp.weixin.qq.com/s?__biz=MzU4MzYxOTIwOQ==&mid=2247499313&idx=1&sn=ca4e74b19cc653f502cf5082326f8519&chksm=fda4f9d7cad370c1ed8413843e090e1968f9353e3ab79d5d396f392a3583e49078633d1e7d71#rd.

———. "Xue Lan and Liang Zheng Interviewed by 'Outlook Weekly' | Is AI Likely to Have Autonomous Consciousness? (薛澜、梁正接受《瞭望》采访 | 人工智能可能有自主意识了吗？)." Weixin Official Accounts Platform, August 15, 2022. http://mp.weixin.qq.com/s?__biz=MzU4MzYxOTIwOQ==&mid=2247492951&idx=1&sn=eef4d58d97c8999178eeb38ed153ed1c&chksm=fda4e2b1cad36ba78c9c4cc2becbc10cc7d9e76389e66d6463291491aa51e996a22cb8340534#rd.

Tsinghua University Foundation Models (THU基础模型). "'About Tsinghua University Foundation Models Research Center' (【清华大学基础模型研究中心简介】)." Weixin Official Accounts Platform, September 25, 2023. http://mp.weixin.qq.com/s?__biz=MzkwMzU1MDMzOQ==&mid=2247484067&idx=1&sn=49cc17fba9ab77814b0adf680a917685&chksm=c095c64ff7e24f59b3f2fd3cb475ddf5dc2222cf32e42f2135e6b5bed421779ec43162d8141b#rd.

———. "SafetyBench: Evaluating the Security of Large Language Models through Multiple Choice Questions (SafetyBench：通过单选题评估大型语言模型安全性)." Weixin Official Accounts Platform, September 20, 2023. http://mp.weixin.qq.com/s?__biz=MzkwMzU1MDMzOQ==&mid=2247484013&idx=1&sn=05dfb8ea209ae5063d00037e25bb8cbd&chksm=c095c681f7e24f9745ee230d5a4c369a65aad9f7d11b3ed9a804fcc6beed25442ddc964f665b#rd.

Tsinghua University (清华大学). "Tsinghua University Establishes a Science and Technology Ethics Committee (清华大学科技伦理委员会成立)," December 31, 2022. https://www.tsinghua.edu.cn/info/1177/100966.htm.

UNESCO. “Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library.” UNESDOC Digital Library, November 23, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

United Nations Office at Geneva. “The Position Paper Submitted by the Chinese Delegation to CCW 5th Review Conference.” Accessed October 12, 2023. https://web.archive.org/web/20190527074927/https://www.unog.ch/80256EDD006B8954/(httpAssets)/DD1551E60648CEBBC125808A005954FA/$file/China%27s+Position+Paper.pdf.

United Nations Security Council. “Security Council Seventy-Eighth Year 9381st Meeting Tuesday, 18 July 2023, 10 a.m. New York, S/PV.9381.” United Nations, July 18, 2023. https://documents-dds-ny.un.org/doc/UNDOC/PRO/N23/210/49/PDF/N2321049.pdf?OpenElement.

Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. “Dual Use of Artificial-Intelligence-Powered Drug Discovery.” *Nature Machine Intelligence* 4, no. 3 (March 7, 2022): 189–91. https://doi.org/10.1038/s42256-022-00465-9.

Vincent, James. “China Is about to Overtake America in AI Research.” The Verge, March 14, 2019. https://www.theverge.com/2019/3/14/18265230/china-is-about-to-overtake-america-in-ai-research.

Vincent Manancourt, Tom Bristow, and Laurie Clarke. “China Expected at UK AI Summit despite Pushback from Allies.” *POLITICO* (blog), August 25, 2023. https://www.politico.eu/article/china-likely-at-uk-ai-summit-despite-pushback-from-allies/.

WAIC. “2019 World Artificial Intelligence Conference (2019年世界人工智能大会).” Accessed October 12, 2023. https://waic2019.sensetime.com/.

WAIC. “2021,” July 31, 2021. https://www.worldaic.com.cn/wangjie?year=2021.

WAIC - World Artificial Intelligence Conference. “A Brief History of WAIC - A Resilient Journey towards Excellence.” LinkedIn, August 12, 2022.

https://www.linkedin.com/pulse/brief-history-waic-resilient-journey-towards-excellence--1c/?trk=organization_guest_main-feed-card_feed-article-content.

Wang, Binjie, Ethan Chern, and Pengfei Liu. “ChineseFactEval: A Factuality Benchmark for Chinese LLMs,” 2023. https://gair-nlp.github.io/ChineseFactEval/.

Wang, Boxin, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. “Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models.” arXiv, January 10, 2022. https://doi.org/10.48550/arXiv.2111.02840.

Wang, Jindong, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, et al. “On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective.” arXiv, August 29, 2023. http://arxiv.org/abs/2302.12095.

Wang Ying. “Record Crowd Expected for WAIC.” China Daily, June 29, 2023. https://www.chinadaily.com.cn/a/202306/29/WS649d557ea310bf8a75d6c5f9.html.

Wang, Yufei, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. “Aligning Large Language Models with Human: A Survey.” arXiv, July 24, 2023. https://doi.org/10.48550/arXiv.2307.12966.

Wang, Zekun, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, et al. “Interactive Natural Language Processing.” arXiv, May 22, 2023. http://arxiv.org/abs/2305.13246.

Weiwen Duan (段伟文). “Chinese Social Science Daily | Duan Weiwen: New Trends in Humanities Reflections on AI (《中国社科报》| 段伟文：人工智能人文反思新动向).” Weixin Official Accounts Platform, September 1, 2020. http://mp.weixin.qq.com/s?__biz=MzU2NDg2OTU0Mg==&mid=2247487148&idx=3&sn=d33cc28ece04d77d641b3958109686ff&chksm=fc452e65cb32a77386b307a19a589233be2ed445b0d66215e133e9120706c860b83ec30140e3#rd.

———. “Duan Weiwen | Accurately Studying the Social and Ethical Risks of Generative AI (段伟文 | 准确研判生成式人工智能的社会伦理风险).” Weixin Official Accounts Platform, May 9, 2023. http://mp.weixin.qq.com/s?__biz=Mzg4NDA5MDEwMw==&mid=2247497715&idx=1&sn=a017479460024c536d0b22b17d2035f9&chksm=cfbfc26bf8c84b7de7b6a86f9c3806d225e6c2fd5d76633c34916bb3919ff0795e6960051cae#rd.

———. “Duan Weiwen | Beyond the Prometheus Difference, Movin towards Robust AI (段伟文｜超越普罗米修斯差异，走向稳健人工智能).” Weixin Official Accounts Platform, December 7, 2021. http://mp.weixin.qq.com/s?__biz=Mzg5OTY0MTc4MA==&mid=2247483903&idx=1&sn=e8ef98e1a57bd8071e5630999319e2df&chksm=c0517924f726f032fb75f477997bc08c8ebba1afa226f7b1c9985f8953f837ff347b132a6597#rd.

———. “Ethical and Political Examination of Algorithm Cognition in the Era of Deep Intelligence (深度智能化时代算法认知的伦理与政治审视).” 中国人民大学学报 36, no. 3 (2022): 23.

———. “[Intelligence and Law] Duan Weiwen: Ethical Strategies as We Approach the Age of AI (【智能与法】段伟文：面向人工智能时代的伦理策略).” Weixin Official Accounts Platform, May 27, 2019. http://mp.weixin.qq.com/s?__biz=MzU0MTU5Nzk5Mg==&mid=2247485593&idx=1&sn=f0b97897f70f827f6267fc7be92d1dbf&chksm=fb26c0decc5149c8b14a89934bc42197df2508104cd6a1d03f8eac7fc256dcb298f9232516ee#rd.

———. “Thousands of Experts Make an Appeal, Can We Press the Pause Button on ChatGPT Research (千名专家呼吁，能让ChatGPT研发按下暂停键吗).” Sina Finance, March 30, 2023. https://finance.sina.cn/tech/2023-03-30/detail-imynrnhk4753471.d.html.

Weixin Official Accounts Platform. “Academic | Guidelines for Artificial Intelligence Safety/Security and Rule of Law (2019) (学界 | 人工智能安全与法治导则)（2019）,” September 5, 2019. https://web.archive.org/web/20231016012639/https://mp.weixin.qq.com/s?src=11&timestamp=1697419307&ver=4837&signature=Vntqb2twpVJwW2Lsh-geszPcx2XzSm27imduy-vj2jgy3IlXI%2AxXGiGrud1XdS8O-VkXS%2AUjqr8wWADTjCjpIOgnoyIFo0d-IjaEcKvo1goRo9fuJhzE-xV5e4is1GPY&new=1.

Weixin Official Accounts Platform. “The ‘Progress and Response Report of the General Large Models (Version 2.0)’ Has Been Officially Released, and Experts Are Discussing ‘Knowledge Production and Political Order in the GPT Era’ (‘通用大模型的进展与应对报告（2.0版）’正式发布 专家热议‘GPT时代的知识生产与政治秩序’),” July 10, 2023. http://mp.weixin.qq.com/s?__biz=MzAxNzAyMzEzNQ==&mid=2649732575&idx=1&sn=bbcf35d4bcd353ed7ad72046964b5626&chksm=83f0fc2fb48775397c0c3d8a9f65e3fd6f8cfc73f1093f85f7e87cb9ae1ec56cf56aea8d5979#rd.

Wen Gao (高文), trans. “Wen GAO.” Chinese Perspectives on AI. Accessed October 12, 2023. https://chineseperspectives.ai/Wen-Gao.

White House. “Ensuring Safe, Secure, and Trustworthy AI,” July 2023. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

White House. “Voluntary AI Commitments,” September 2023. https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

World Economic Forum. “Build an ethical "ark" to make artificial intelligence knowable, controllable, available and reliable (打造伦理“方舟”，让人工智能可知、可控、可用、可靠),” December 27, 2018. https://cn.weforum.org/agenda/2018/12/b9b7b1dc-5021-4f7d-965b-f4cf3bb4cb3b/.

Wu, Baoyuan, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. “BackdoorBench: A Comprehensive Benchmark of Backdoor Learning,” 2022. https://openreview.net/forum?id=31_U7n18gM7.

Wu, Baoyuan, Li Liu, Zihao Zhu, Qingshan Liu, Zhaofeng He, and Siwei Lyu. “Adversarial Machine Learning: A Systematic Survey of Backdoor Attack, Weight Attack and Adversarial Example.” arXiv, February 18, 2023. http://arxiv.org/abs/2302.09457.

Xi Jinping (习近平). “Full Text: Remarks by Chinese President Xi Jinping at the 15th BRICS Summit.” The State Council, August 24, 2023.

https://english.www.gov.cn/news/202308/24/content_WS64e693a0c6d0868f4e8dec78.html.

Xiaoyuan Jiang (江晓原). “Jiang Xiaoyuan | Why Is It Inevitable That AI Will Threaten Our Civilization? (江晓原 | 为什么人工智能必将威胁我们的文明？).” Weixin Official Accounts Platform, August 2, 2016. http://mp.weixin.qq.com/s?__biz=MjM5MjE2MzY1OA==&mid=2674934092&idx=1&sn=55fa6d5eeb9fd5fbb234b2a634e3f7c9&chksm=bc2e93698b591a7f409edc2c900b4da222ed3e8e2406945c76629618bb14b9dfe3662568fe0b#rd.

———. “Scientific and Technological Innovation Should Be Established upon Bottom-Line Thinking - Taking the Development of AI as an Example (科技创新应树立底线思维——以人工智能发展为例).” CPC News, July 29, 2016. http://theory.people.com.cn/n1/2016/0729/c40531-28593493.html.

Xinhua. “Ant Group Has Establishes a Science and Technology Ethics Advisory Committee, Consisting of 7 Experts in the Fields of Science and Technology and Social Sciences (蚂蚁集团成立科技伦理顾问委员会，由7位科技及社会科学领域专家构成),” March 6, 2023. https://www.xinhuanet.com/tech/20230306/24e6bf97479d4eae8be12bf76d9b812b/c.html.

Xinhua. “Hawking Suggests AI Threat Theory Again: Might Cause Human Extinction (霍金再抛人工智能威胁论：或招致人类灭亡),” April 28, 2017. https://web.archive.org/web/20190107075125/http://www.xinhuanet.com/tech/2017-04/28/c_1120889914.htm.

Xinhua. “‘Iron Man’ Elon Musk Suggests AI Threat Theory Again (钢铁侠”马斯克再抛人工智能威胁论),” September 7, 2017. https://www.xinhuanet.com/world/2017-09/07/c_129697709.htm.

Xinhuanet. “Xi Jinping: Promote the Healthy Development of China’s New Generation AI (习近平：推动我国新一代人工智能健康发展),” October 31, 2018. https://www.xinhuanet.com/politics/leaders/2018-10/31/c_1123643321.htm.

Xinhuanet. “Xi Jinping’s Speech at the 15th BRICS Leaders Summit - Full Document (习近平在金砖国家领导人第十五次会晤上的讲话 - 全文),” August 23, 2023. https://www.news.cn/politics/leaders/2023-08/23/c_1129819257.htm.

Xiuquan Li (李修全). *The Intelligent Revolution: The Evolution and Value Creation of AI Technology (智能化变革: 人工智能技术进化与价值创造)*. Tsinghua University Press (清华大学出版社), 2021. https://www.amazon.com/%E6%99%BA%E8%83%BD%E5%8C%96%E5%8F%98%E9%9D%A9-%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E6%8A%80%E6%9C%AF%E8%BF%9B%E5%8C%96%E4%B8%8E%E4%BB%B7%E5%80%BC%E5%88%9B%E9%80%A0-%E6%9D%8E%E4%BF%AE%E5%85%A8/dp/7302578443.

Xiuquan Li (李修全). “Xiuquan LI.” Translated by Concordia AI. Chinese Perspectives on AI. Accessed October 12, 2023. https://chineseperspectives.ai/Xiuquan-LI.

Xu, Guohai, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, et al. “CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility.” arXiv, July 18, 2023. http://arxiv.org/abs/2307.09705.

Xu, Liang, Kangkang Zhao, Lei Zhu, and Hang Xue. “SC-Safety: A Multi-Round Open-Ended Question Adversarial Safety Benchmark for Large Language Models in Chinese.” arXiv, October 9, 2023. https://doi.org/10.48550/arXiv.2310.05818.

Xu Wei. “Xi to Address Opening of Zhongguancun Forum.” China Daily, September 4, 2021. https://www.chinadaily.com.cn/a/202109/24/WS614d0374a310cdd39bc6b221.html.

Xuanjing Huang (黄萱菁). “About Me.” Xuanjing Huang. Accessed October 12, 2023. https://xuanjing-huang.github.io/.

Yan Yu (喻琰). “The Bund Conference | Ant Group Launches the Ethical Co-Construction Plan for Large Models, Giving Legal Exams to the Large Model (外滩大会丨蚂蚁集团启动大模型伦理共建计划，给大模型出法律

考题).” The Paper (澎湃), September 9, 2023. https://m.thepaper.cn/newsDetail_forward_24548122.

Yanyong Du (杜严勇). “Security of Artificial Intelligence: Problems and Solutions (人工智能安全问题及其解决进路).” *Philosophical Trends (哲学动态)*, no. 9 (2016): 99–104.

———. “Security of Artificial Intelligence:Problems and Solutions (杜严勇：人工智能安全问题及其解决进路).” China Big Data Industry Observatory (中国大数据产业观察), February 27, 2017. http://www.cbdio.com/BigData/2017-02/27/content_5459010.htm.

Yao, Jing, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. “From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models.” arXiv, September 3, 2023. http://arxiv.org/abs/2308.12014.

Yi Wang (王毅). “Staying Open and Inclusive and Upholding Multilateralism: Toward a Community with a Shared Future for Mankind.” Ministry of Foreign Affairs, May 26, 2021. https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/zyjh_665391/202105/t20210526_9170548.html.

Yi Zeng (曾毅). “Opinion on Strengthening the Ethics and Governance in Science and Technology [China].” International Research Center for AI Ethics and Governance, March 22, 2022. https://ai-ethics-and-governance.institute/2022/03/22/china-released-opinion-on-strengthening-the-ethics-and-governance-in-science-and-technology/.

Yi Zeng (曾毅) and Kang Sun. “Whether We Can and Should Develop Strong AI: A Survey in China.” Center for Long-term Artificial Intelligence. Center for Long-term Artificial Intelligence, March 12, 2023. https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence.

Yi Zeng (曾毅), Sun Kang, Lu Enmeng, and Zhao Feifei. “Voices from China on ‘Pause Giant AI Experiments: An Open Letter.’” Center for Long-term Artificial Intelligence, April 4, 2023.

https://long-term-ai.center/research/f/voices-from-china-on-pause-giant-ai-experiments-an-open-letter.

Yicai (第一财经). “Exclusive Dialogue with Ma Yi, Dean of Mathematics at Hong Kong University: Concern about AI’s Domination of the World Is Unfounded and the ‘Singularity’ Is Far from Coming (独家对话港大数科院长马毅：担心AI统治世界是杞人忧天，‘奇点’远未到来).” Sina Mobile, April 20, 2023. https://finance.sina.cn/2023-04-20/detail-imyqzcvt5065964.d.html?from=wap.

Yidong Liu (刘益东). “Comparative Analysis and Evaluation of Two Types of Science and Technology Ethics (对两种科技伦理的对比分析与研判).” People’s Tribune (人民论坛), June 2, 2022. https://web.archive.org/web/20220705220542/http://www.rmlt.com.cn/2022/0602/648400.shtml.

Yiming Cui. “Chinese-LLaMA-Alpaca/README_EN.Md at Main · Ymcui/Chinese-LLaMA-Alpaca.” GitHub, April 17, 2023. https://github.com/ymcui/Chinese-LLaMA-Alpaca.

Yin, Zhangyue, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. “Do Large Language Models Know What They Don’t Know?” arXiv, May 30, 2023. http://arxiv.org/abs/2305.18153.

Yu, Yaodong, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. “White-Box Transformers via Sparse Rate Reduction.” arXiv, June 1, 2023. https://doi.org/10.48550/arXiv.2306.01129.

Yu, Yaodong, Tianzhe Chu, Shengbang Tong, Ziyang Wu, Druv Pai, Sam Buchanan, and Yi Ma. “Emergence of Segmentation with Minimalistic White-Box Transformers.” arXiv, August 30, 2023. https://doi.org/10.48550/arXiv.2308.16271.

Yuan, Luyao, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. “In Situ Bidirectional Human-Robot Value Alignment.” *Science Robotics* 7, no. 68 (July 13, 2022): eabm4183. https://doi.org/10.1126/scirobotics.abm4183.

Yuan, Zheng, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. "RRHF: Rank Responses to Align Language Models with Human Feedback without Tears." arXiv.org, April 11, 2023. https://arxiv.org/abs/2304.05302v3.

Zhang, Ge, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, et al. "Chinese Open Instruction Generalist: A Preliminary Release." arXiv, April 24, 2023. http://arxiv.org/abs/2304.07987.

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu. "Interpretable Convolutional Neural Networks." arXiv, February 14, 2018. https://doi.org/10.48550/arXiv.1710.00935.

Zhang, Quanshi, and Song-Chun Zhu. "Visual Interpretability for Deep Learning: A Survey." arXiv, February 7, 2018. http://arxiv.org/abs/1802.00614.

Zhang, Zhaowei, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. "Heterogeneous Value Evaluation for Large Language Models." arXiv, June 1, 2023. http://arxiv.org/abs/2305.17147.

Zhang, Zhengyan, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, et al. "CPM: A Large-Scale Generative Chinese Pre-Trained Language Model." arXiv, December 1, 2020. http://arxiv.org/abs/2012.00413.

Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. "SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions." arXiv, September 13, 2023. http://arxiv.org/abs/2309.07045.

Zhaoxu Ma (马朝旭). "Statement by Ambassador Ma Zhaoxu At the United Nations High-Level Conference of Heads of Counter-Terrorism Agencies of Member States," June 28, 2018. https://www.un.org/counterterrorism/ctitf/sites/www.un.org.counterterrorism.ctitf/files/S1-China.pdf.

Zhihua Zhou (周志华). "Article: On Strong Artificial Intelligence (Zhou Zhihua)." Translated by Jeffrey Ding. Google Docs, 2018. https://docs.google.com/document/d/1RP_bWfC1waWQaLwunQBN_R0yRNIDjVOOE4rhmqm8JSA/edit.

Zhijian Xia (夏志坚). "The AI Ethical Governance Challenge: What Do Experts from China, America, Europe, Japan, and the UK Think? (AI的伦理治理挑战：中美欧日英各方专家怎么看？)." Caixin (财新), November 1, 2019. https://zhishifenzi.blog.caixin.com/archives/214934.

Zhou, Bolei, David Bau, Aude Oliva, and Antonio Torralba. "Comparing the Interpretability of Deep Networks via Network Dissection." In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 243–52. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-28954-6_12.

Zhou, Jingyan, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. "Towards Identifying Social Bias in Dialog Systems: Frame, Datasets, and Benchmarks." arXiv, October 28, 2022. http://arxiv.org/abs/2202.08011.

Zhou, Jingyan, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. "Rethinking Machine Ethics -- Can LLMs Perform Moral Reasoning through the Lens of Moral Theories?" arXiv, August 29, 2023. http://arxiv.org/abs/2308.15399.

Zhu, Junhua. "AI Ethics with Chinese Characteristics? Concerns and Preferred Solutions in Chinese Academia." *AI & SOCIETY*, October 17, 2022. https://doi.org/10.1007/s00146-022-01578-w.

Zhu, Kaijie, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, et al. "PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts." arXiv, August 24, 2023. http://arxiv.org/abs/2306.04528.

ZTE. "ZTE Showcases Data-Driven Intelligence at ITU AI for Good 2023 Summit," July 19, 2023. https://www.zte.com.cn/content/zte-site/www-zte-com-cn/global/about/news/zte-showcases-data-driven-intelligence-at-itu-ai-for-good-2023-summit.

CONCORDIA AI
安远AI