CENTER
FOR
DATA
INNOVATION

# How Experts in China and the United Kingdom View AI Risks and Collaboration

By Yanzi Xu and Daniel Castro | August 12, 2024

**As AI continues to advance, the technology has created many opportunities and risks. Despite significant geopolitical differences, a series of interviews with AI experts in China and the United Kingdom reveals common AI safety priorities, shared understanding of the benefits and risks of open source AI, and agreement on the merits of closer collaboration—but also obstacles to closer partnerships. Fostering a closer relationship could help both countries achieve their objectives of developing innovative, safe, and reliable AI.**

## INTRODUCTION

Recent advances in artificial intelligence (AI) offer many important benefits for society, the economy, and scientific progress. One important factor in these advancements is the development of open source AI: AI technologies whose source code and data are freely available for others to use, study, modify, and distribute.[1] Open source AI is crucial for accelerating innovation through collaborative development, reducing redundancy, and democratizing access to AI capabilities. Since it is available to anyone, it facilitates economic development and social progress, such as by allowing anyone to adapt and fine-tune highly capable AI models for specific tasks. Moreover, global collaboration among researchers, developers, and users on open source AI enables collective progress on shared AI projects and promotes the development of guidelines and best practices for transparency, accountability, and ethics. Finally, by fostering transparency and accountability through accessible code and data, it helps identify and address biases, errors, and ethical concerns in AI and allows users to understand how the technology works.

Both China and the United Kingdom are active members of the open source community and pioneers in the field of AI. This history has led them to emerge as leaders in open source AI. For example, the U.K.-based company Stability AI is the developer of many popular open source generative AI tools used for creating images, audio, 3D models, and code. And China has produced some of the top-performing open source large language models (LLMs) in the world, including Qwen (Alibaba) and Yi (01.AI).[2] These open source AI projects provide competition to proprietary (or "closed") AI where developers restrict public access to the underlying technology.

However, open source AI presents unique challenges. First, unlike proprietary AI, where developers can provide oversight of what users do with their technology, once developers make open source AI publicly available, they have little to no control over how others might use their technology. Malicious actors may tamper with open source AI to remove safeguards, manipulate results, and generate inaccurate information. In addition, malicious actors may use the technology for dangerous and illicit purposes, such as to conduct cyberattacks, spread disinformation, commit fraud, create contraband, and engage in other illegal activities. Second, unlike proprietary AI, where the developer is responsible for the technology, there is not always someone responsible for open source AI projects. As a result, the technology may have known bugs or security vulnerabilities that nobody addresses. Similarly, open source AI may be provided without any warranties or guarantees. For example, users may not know if developers trained an open source AI model on poor-quality or illicit data. Finally, the development practices of open source products can create unknown risks, such as if attackers surreptitiously attempt to introduce malicious code or data into an open source project.

Addressing risks from AI is an issue of global concern, and one at which both the United Kingdom and China have remained at the forefront, even as they both seek to support their respective firms' development and use of AI. The United Kingdom convened an AI Safety Summit in 2023, which many countries attended, including China. The summit concluded with the Bletchley Declaration whereby participating countries resolved "to sustain an inclusive global dialogue that engages existing international fora and other relevant initiatives and contributes in an open manner to broader international discussions."[3] President Xi Jinping later reiterated this call for mutually beneficial cooperation on common interests, including AI, in remarks in San Francisco at a bilateral meeting with the U.S. president.[4] The Chinese Ministry of Foreign Affairs also released a statement in October 2023 that calls for "global collaboration to foster the sound development of AI, share AI knowledge, and make AI technologies available to the public under open source terms."[5]

Despite these high-level government declarations, it is unclear whether the United Kingdom and China can turn their aspirations for closer cooperation in AI into meaningful action. To understand the feasibility of such partnerships, it is important to better understand both whether the concerns and priorities of AI experts outside government align and their experiences to date on collaboration. This report strives to provide insights into these issues.

## METHODOLOGY

This study employed a qualitative research design to explore the perspectives of experts on the risks associated with AI, especially open source AI, where there are potential pathways for collaboration. Qualitative methods are particularly well-suited for capturing the depth and complexity of expert opinions, providing rich, detailed data that can inform understanding of a rapidly evolving field.

From March to June 2024, we conducted in-depth interviews with a purposive sample of experts from academia and industry in the United Kingdom and China. We selected participants based on their expertise in AI technology and policy, ensuring a diverse range of perspectives. The final sample included 24 experts (19 from China; 5 from the United Kingdom) from universities, research institutes, technology companies, and consulting agencies. We primarily focused on China-based experts for these interviews, given that views of U.K.-based experts are already widely documented in Western media.

We collected data through semi-structured interviews, which allowed for flexibility in exploring various topics while ensuring that all key areas of interest were covered. We conducted interviews either in person or via online meetings, depending on the location and availability of the participants. Based on participants' requests, the interviews were conducted in one of two formats, which participants could choose based on their preferences: a one-time recorded interview or multiple short conversations not recorded. Regardless of the format, each complete interview lasted around one hour, and the recorded ones got the participants' consent for transcription and analysis; we took handwritten notes for the unrecorded interviews. We reviewed and edited transcriptions for clarity and completeness, and removed personal identifiers to maintain participant confidentiality.

The analysis followed a systematic process to identify patterns within the collected data and was conducted in several stages:

1. **Initial Coding:** We began by reading through the transcripts and notes multiple times to become familiar with the data, and then developed an initial coding framework based on recurring topics and ideas. This involved labeling relevant sections of the text with descriptive codes.

2. **Codebook Development:** We developed a comprehensive codebook to guide the coding process. The codebook included definitions for each code, along with examples from the data to illustrate their application. We organized codes into broader themes and sub-themes to capture the complexity of the data.

3. **Transcript Coding:** Using the codebook, we systematically coded the transcripts. Discrepancies in coding were discussed and resolved through consensus.

4. **Theme Identification:** After coding, we reviewed the data to identify overarching themes and patterns. Themes were derived by grouping related codes and examining the relationships between them. This process involved iterative review and refinement to ensure that the themes accurately represented the data.

5. **Synthesis:** The final themes were interpreted in the context of the research questions. This involved synthesizing the findings to provide a coherent narrative that addressed the study's aims. Patterns and insights were highlighted to illustrate the key perspectives of the participants.

## FINDINGS AND DISCUSSION

The purpose of this study was to provide a detailed understanding of the perspectives of U.K. and China experts to build active connections. The findings offer valuable insights into the key concerns, priorities, and potential areas for collaboration in AI safety. Key themes that emerged from the analysis were AI safety priorities, benefits and risks of open model weights, AI regulations, and international collaboration barriers. We translated and edited quotations mentioned in the following sections for clarity. We also assigned interviewees a random number with a code—UK for the United Kingdom, CN for China—to maintain confidentiality.

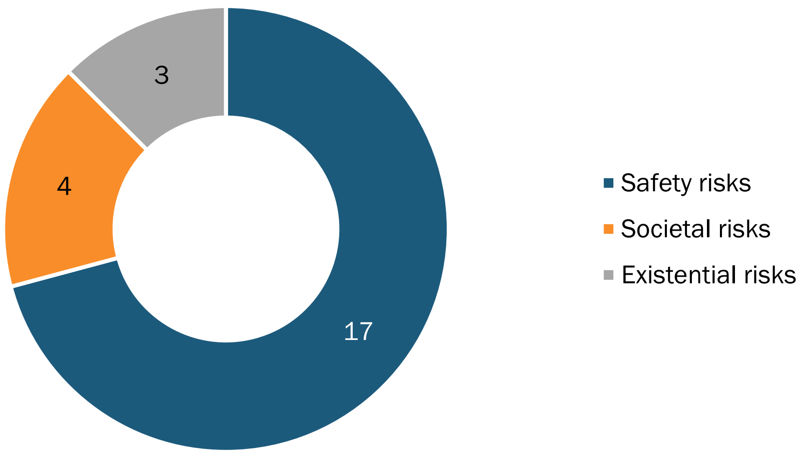
### Top Priorities for AI Safety

Interviewed experts shared their top concerns about AI safety. To capture these concerns, only their initial responses to the question, "What worries you most about AI safety?" were coded under this theme. Although all experts discussed multiple AI issues and the interviews covered various subtopics, the data may still show no or limited concern in specific areas. This does not imply that the experts are indifferent to those areas, only that it was not their top priority.

Experts mentioned risks in three categories: safety risks (i.e., risks of creating unsafe AI), societal risks (i.e., risks of negative impact on society from AI), and existential risks (i.e., safety or societal risks from AI that could create an irreversible global catastrophe). Both U.K. and Chinese experts highlighted several risks associated with unsafe AI, such as AI systems that are inaccurate, unreliable, or fail under unexpected conditions. They also expressed concerns about the safety of human-AI interactions, such as a robot harming a user or self-driving car accidents, and the risk of building AI systems that do not have human goals and values correctly encoded.

Experts also listed various societal risks, including misuse of AI, such as using AI to create deepfakes used for fraud. U.K. experts were more concerned with other societal risks such as the ethical implications of using AI with children and the potential for AI to replace human workers and thus lead to a high unemployment rate. They stressed the importance of addressing these ethical concerns to prevent social problems. In contrast, some Chinese experts focused on the potential existential threats of advanced AI systems and AI weaponization. They highlighted the broader implications for humanity, recognizing that AI's rapid development could lead to significant global changes.

**Figure 1: Top concerns in AI safety for the United Kingdom and China (# of respondents)**



## Benefits and Risks of Open Source AI

This report uses the term "open source AI," even though some AI experts in the field dispute the use of this term.[6] In fact, most AI systems exist along some spectrum of "open" and "closed" depending on which elements and artifacts developers have made public, such as code, data, and documentation.[7] For example, some developers of "open source AI" may not release the training data or code used to develop the system, and instead only release the model architecture and parameters that allow for others to reuse the AI model.

All interviewed experts agreed that open source AI models offer significant benefits. Compared with closed models, they believe open models can undergo more thorough inspections to improve quality, promote collaboration within open source communities, build expert networks, and provide independent developers access to models trained on large datasets they otherwise could not collect.

Regarding risks, experts consistently identified malicious use as their top concern. They also raised concerns about data privacy violations, which arise from the improper handling and collecting of private information, potentially leading to unauthorized access and misuse of personal data. They agreed that, regardless of whether a model is open or closed, individuals with bad intentions can attack the model itself to circumvent safeguards or manipulate outputs, or use it to cause harm.

> “[I]f people are going to do bad things with AI, it doesn’t really matter if it's closed source or open source, they’re still going to find a way to do it.” — 2-UK
>
> “If someone wants to attack a model and has the knowledge and technology to do so, whether the model is open or closed does not matter.” — 3-CN

These perspectives align with expert views in other countries assessing the risks and benefits of open source AI. For example, multiple studies have identified potential avenues for misuse of open source AI, including to produce targeted phishing attacks, disinformation, biosecurity threats, voice cloning, nonconsensual intimate imagery, and child sexual abuse material.[8] However, the marginal risk from misuse of open source AI models—the extent to which open source models introduce additional risk of misuse compared with closed models or other existing technologies such as the Internet—is often low.[9]

Interviewees noted that preventing different attacks presents a complex challenge that would likely require ongoing monitoring and regular updates to safety countermeasures. Additionally, some expressed concern about being underprepared for frontier AI developments, as rapid advancements may outpace the establishment of safety and ethical guidelines, leaving gaps in readiness to address emerging challenges. Furthermore, some believe the lack of standards to guide the use, development, and management of open source AI complicates the implementation of consistent practices, increasing the risk of misuse and hindering collaborative efforts to ensure AI safety and reliability.

## AI Regulations

Experts from both nations believe that their countries’ approaches toward AI regulation are reasonable given current circumstances. In the United Kingdom, the government outlined its vision for a “pro-innovation approach to AI regulation” in a white paper in August 2023.[10] In China, the government is considering the “Model Artificial Intelligence Law (MAIL) v.2.0,” a proposed law for regulating the developers and providers of AI.[11]

While acknowledging room for policy refinement and the need for the government to articulate clearer details as it moves forward, experts found that the general trajectory aligns well with each nation's specific domestic priorities, such as taking a sectoral-based approach to regulation in the United Kingdom and promoting open source AI in China.

"I think the U.K. government has taken a fairly reasonable and measured approach. They have taken their time, emphasized the application of sectoral regulation to AI uses. They kicked off a search for gaps in regulation. They empowered sector regulators. This is good." — 3-UK

"The draft [MAIL v.2.0] needs more details and clarifications on some terms it used. Since it's a draft, it can't lead to actual changes, but it sets a positive direction for the future policies. We can see that it promotes the development of open source AI via financial support and tax deduction in relevant fields, which are good." — 2-CN

On the topic of increasing transparency in AI, many experts believe there is a significant knowledge gap between AI developers and the general public, which limits how much developers can do on their own to build public trust by disclosing additional information about the development of their AI products. However, some experts noted that greater transparency from developers could be beneficial if independent third-party organizations could evaluate this information and produce accessible reports to assist public understanding.

Several experts also shared their views on the future development of AI policies, although opinions varied. Some argued for greater regulation to guide the development of AI products, whereas others urged regulators to improve their capacity to evaluate and monitor risks from the development and use of AI.

"To manage AI risks, we must address their origins by focusing on developers. There should be regulations to guide the development of AI products to ensure responsible development. For example, there should be consequences if developers train models with biased data or ignore safety protocols when releasing AI products." — 7-CN

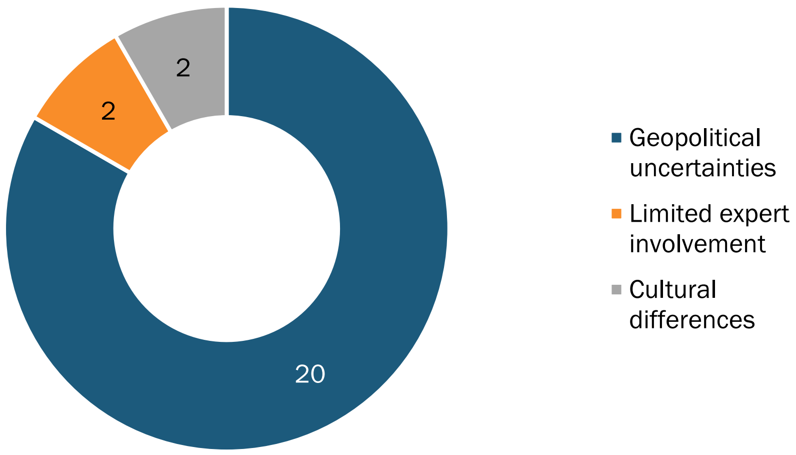
"So upskilling regulators, focusing on the specific use cases rather than regulating the technology itself, and then things that we have done for a long time, creating new opportunities and regulatory incentives to support innovation and usage, and support experimentation." — 2-UK

"Legislators should establish an evaluation framework to manage AI risks effectively. To prevent uncontrollable scenarios caused by existential risks, they should design identifiers for each development stage to assess the model's strength and ensure AI technology remains under control." — 4-CN

## International Collaboration on AI

International collaboration is seen as crucial to effectively address AI risks. As the U.S. Department of Commerce wrote in a recent report, "The safe, secure, and trustworthy deployment of AI requires coordination with allies and partners."[12] Both U.K. and Chinese experts emphasized the importance of working together to develop robust standards and governance mechanisms. However, they also identified several barriers to collaboration, including geopolitical uncertainties, cultural differences, and limited involvement of content experts.

**Figure 2: Barriers to international collaboration (# of respondents)**



Most experts from both nations consider geopolitical uncertainties the most significant factor hindering international collaboration, and it is hard to ignore the influences of geopolitics. Interviews reveal that trust issues negatively impact scholars' readiness to engage in communication, and unpredictable international relations, especially between the United States and China, profoundly affect collaboration.

"It is really hard to negotiate when your partner is saying one thing and doing something completely different. As much as we would like to talk about AI and pretend like it's not connected to these broader geopolitical issues, it absolutely is because this is the future."
— 2-UK

"Competitions between China and the U.S. can lead to issues, including mutual antagonism and hostility, and sometimes AI is used to create divisions and misunderstandings among people from different countries, which is something I really don’t want to see.” — 1-CN

“Another big issue is the impact of geopolitics, both in terms of the dynamic between the U.S. and China and the fragmentation of multiple countries, which can be seen in export controls and the entire AI chip industry … This worries me the most because there’s already a baseline risk between the U.S., China and the potential for a third World War-type scenario.” — 1-UK

Experts also mentioned cultural differences as factors that could hinder collaboration. Variations that exist between people from different backgrounds, such as values, beliefs, communication styles, and social norms, can influence how individuals perceive and interact with the world. In the context of international collaboration on AI safety, differently trained AI models, domestic regulations, and varying interests can make it more challenging for countries to reach an agreement on AI cooperation.

Some experts who focus on developing AI technologies believe that the lack of other technical experts involved in international dialogues means that international collaborations will not lead to practical solutions and demotivates them from participating in future activities. In addition, there is not a single, consistent view among a country’s experts on AI safety issues. Finally, experts noted that language barriers can create a bubble between different expert groups and prevent collaboration and knowledge-sharing.

Several U.K. experts believe a third country outside the U.S.-China competition, such as the United Kingdom, should take the lead to promote international collaboration. On the other hand, all Chinese experts we spoke with believe that the United Nations, or another global institution that will not exclude any country, should take the lead. Instead of regarding the United Kingdom as a suitable “middleman” in collaboration, Chinese experts tend to group the United Kingdom together with the United States as Western countries.

“Given the political climate of the U.S. and China going into this election in the fall, we have seen a breathing period where there has been an opening in space for dialogue … Countries like the U.K. and beyond will play a big role in creating space and generating dialogue.” — 4-UK

“It appears that the United States prefers to first establish consensus among its allies. However, I think no country should be excluded from international collaboration. If countries all search for allies and form exclusive small circles, there would be no chance to collaborate.” — 5-CN

## CONCLUSION

Experts from the United Kingdom and China share many of the same concerns and priorities about AI, and these commonalities can serve as the basis for future collaboration and cooperation between these countries, especially around open source AI. Further dialogue and joint research on open source AI offers a unique opportunity to identify and monitor emerging risks, develop technical standards and solutions, and evaluate potential coordination on oversight measures. By working together on AI risk management, the United Kingdom and China can make progress toward a shared goal of a safer and more reliable future for open source AI.

To move forward, experts from government, industry, academia, and civil society should continue to develop neutral platforms to foster inclusive discussions about AI opportunities and risks and to build trust for productive collaboration on AI safety. Next steps for these stakeholders should include enhancing cross-cultural understanding through joint research and international exchanges, encouraging a diverse range of stakeholders to participate in global AI safety discussions, and promoting clear and timely communication about AI safety developments in multiple languages. While many geopolitical tensions will persist, by increasing collaboration on open source AI, stakeholders in the United Kingdom and China can promote the responsible development and implementation AI technologies that balance innovation and safety.

## REFERENCES

1. "The Open Source AI Definition – draft v. 0.0.8," Open Source Initiative, https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8.

2. "Open LLM Leaderboard," HuggingFace, n.d., https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard (accessed July 27, 2024).

3. "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," Gov.UK, November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

4. "Xi Jinping holds meeting with US President Biden," *Xinhua News Agency*, November 17, 2023, http://www.news.cn/politics/leaders/2023-11/16/c_1129977979.htm.

5. "Global AI Governance Initiative," Ministry of Foreign Affairs, The People's Republic of China, October 20, 2023, https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html.

6. Elizabeth Gibney, "Not all 'open source' AI models are actually open: here's a ranking," *Nature*, June 19, 2024, https://www.nature.com/articles/d41586-024-02012-5.

7. Matt White et al., "The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI," Linux Foundation, April 17, 2024, https://lfaidata.foundation/blog/2024/04/17/introducing-the-model-openness-framework-promoting-completeness-and-openness-for-reproducibility-transparency-and-usability-in-ai/.

8. Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," arXiv.org, February 27, 2024, https://arxiv.org/pdf/2403.07918v1.

9. Ibid.

10. "A pro-innovation approach to AI regulation," Gov.UK, August 3, 2023, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

11. Hui Zhou et al., "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Zenodo, April 16, 2024, https://zenodo.org/records/10974163.

12. "Dual-Use Foundation Models with Widely Available Model Weights," NTIA, July 2024, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

## ACKNOWLEDGMENT

This report was made possible by generous support from the British Embassy in Beijing. The Center maintains complete editorial independence for all of its work. All opinions, findings, and recommendations are those of the Center and do not necessarily reflect the views of its supporters. Any errors and omissions are the author's responsibility alone.

## ABOUT THE AUTHORS

Yanzi Xu is a research fellow at the Information Technology and Innovation Foundation (ITIF). Prior to joining, she worked as a research assistant and a learning support associate in the field of higher education. Yanzi holds a master's degree in learning sciences and technologies from the University of Pennsylvania and a bachelor's degree in TESOL education from the Ohio State University.

Daniel Castro is the director of the Center for Data Innovation and vice president of ITIF. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

## ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of ITIF, a nonprofit, nonpartisan think tank.

**Contact: info@datainnovation.org**

**datainnovation.org**