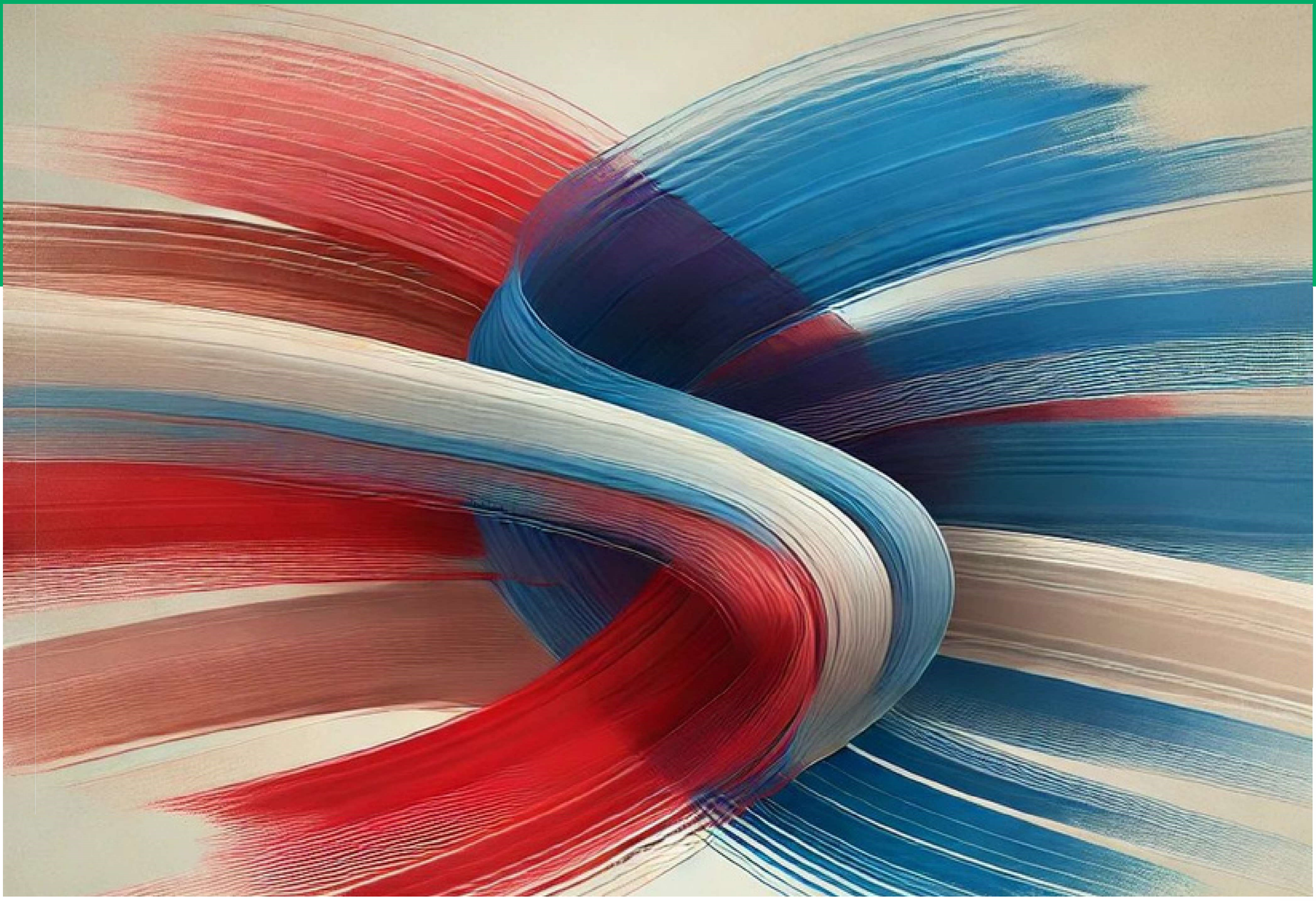
RESEARCH PAPER
JANUARY 2025

AI GOVERNANCE INITIATIVE
OXFORD MARTIN SCHOOL
UNIVERSITY OF OXFORD

Promising Topics for US-China Dialogues on AI Safety and Governance



Authors: Saad Siddiqui, Kristy Loke, Stephen Clare, Marianne Lu, Aris Richardson, Lujain Ibrahim, Conor McGlynn and Jeffrey Ding

In partnership with:

Safe AI Forum

Promising Topics for US–China Dialogues on AI Safety and Governance

Saad Siddiqui[1] Kristy Loke[2] Stephen Clare[3] Marianne Lu[4]

Aris Richardson[3] Lujain Ibrahim[5] Conor McGlynn[6]

Jeffrey Ding[7]

January 2025

The views expressed in this report are also not representative of the institutions the authors are affiliated with or employed by, and should be considered purely personal opinions.

[1] Safe AI Forum
[2] Independent
[3] Centre for the Governance of AI
[4] Stanford University
[5] University of Oxford
[6] Harvard University
[7] George Washington University

Executive Summary

In 2023, both the US and China signed the Bletchley Declaration, acknowledging the potential for serious harm from advanced AI systems as well as the importance of cooperating to mitigate it.[1] Since then, the US and China have also jointly signed UN resolutions on AI issues, held a round of intergovernmental dialogues on AI, and agreed to limit the integration of AI into control systems for nuclear weapons. The number of Track II dialogues has similarly increased.[2]

In this report, we develop recommendations for AI governance and safety dialogue topics from one specific angle: identifying topics on which there is significant common ground between the US and China, through a comprehensive analysis of over 40 key primary AI policy and corporate governance documents from the two countries. We analyze these areas of common ground in the context of US–China relations, setting aside topics that would be deemed too sensitive, or linked to existing tensions between the two countries.

Through this analysis, we arrive at recommendations for dialogue topics and suggestions for which diplomatic tracks (i.e. Track I or Track II) could adopt these recommendations. While the focus of our recommendations is the bilateral context, many of these discussions could also take place in the multilateral context, through fora such as the AI summit series.

Our four key recommendations are:

1. **US and Chinese governments should strengthen existing intergovernmental dialogue, covering issues related to national security, such as evaluating models for 'dangerous capabilities' and preventing proliferation to non-state actors.**[3]

   Both countries emphasize the importance of testing and evaluating AI systems, and are concerned about the dangerous national security-relevant capabilities that AI systems may possess. While there are some differences – Chinese safety/security assessments so far have focused less on dangerous national security-relevant capabilities of AI models and more on control of politically sensitive content – there is common ground and evidence of consensus from Track 2 processes such as the International Dialogues on AI Safety, for the two governments to discuss this topic further.

   The governments could start to agree on the domains in which evaluations are important, eventually moving towards defining common critical thresholds, such as red lines, which if crossed would signal that AI models pose significant domestic and global national security risks.

[1] "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 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

[2] Nayan Chandra Mishra, "From Competition to Cooperation: Can US–China Engagement Overcome Geopolitical Barriers in AI Governance?" *Tech Policy Press* September 2024 https://www.techpolicy.press/from-competition-to-cooperation-can-uschina-engagement-overcome-geopolitical-barriers-in-ai-governance/

[3] By dangerous capabilities here we mean a cluster of AI system capabilities that may have national security implications, such as CBRN, model autonomy, cyberattacks, and persuasion.

2. **The US and Chinese governments and their respective companies should explore technical standards-setting discussions on reliability, robustness and adversarial testing, either in existing standards-setting bodies (e.g. the International Standards Organisation) or new fora.**

   There is keen attention on the need for reliability, robustness and adversarial testing from both the US and China. The scope of these discussions should be restricted to issues unrelated to national security considerations of AI systems, instead focusing largely on commercial product safety. These topics may prove especially salient as AI companies from both the US and China increasingly sell products internationally.

3. **International industry consortia should consider either involving Chinese companies in existing dialogues or setting up distinct tracks of dialogue with Chinese actors, particularly on content provenance.**

   There is common concern about the need for better information traceability and watermarking across both US and Chinese governments and companies. C2PA is an existing industry association that already brings together leading companies in the US to tackle this issue, and could consider working with Chinese companies to build global norms and standards.

4. **Track II dialogues should include discussions on emerging or novel approaches to safety and governance, especially those for which there may be existing common ground between US and Chinese perspectives.**

   There is common ground on two issues in particular. First, there is interest in developing new governance mechanisms for AI. Track 2s such as the Yale Law School and Chinese Academy of Social Science Law Centre's Track II dialogue could tap on in the form of technical, detailed and concrete best practice sharing related to novel approaches to governance such as compute thresholds and model registration systems. There is also common understanding that AI can be used to enhance AI safety, such as by aiding in or even automating evaluations of new models. Scientific Track 2 dialogues, such as International Dialogues on AI Safety could leverage this common ground by discussing how AI systems can be used to improve AI safety.

We believe that the above recommendations represent potential 'low-hanging fruit' in international scientific diplomacy, which could continue to be promising discussion topics even as geopolitical tensions wax and wane. This is not an exhaustive list of all topics relevant for US–China dialogues on AI safety and governance, nor a list of topics that existing dialogues fail to cover. Nor do we consider topics that may need to be discussed despite a lack of existing common ground. Instead, our recommendations are topics for which, we believe, there is common ground between domestic American and Chinese conceptions of AI safety and governance.

Finally, it is also worth acknowledging that while dialogues themselves typically entail minimal risk, some of the topics we have recommended above may not be suitable for deeper cooperation. For example, while the US and Chinese government can try to find common ground on which domains require dangerous capability evaluations, deeper cooperation through joint testing or sharing of specific evaluation methodology may be deemed too risky. We recognise such risks and note that further work is required to identify promising AI safety and governance issues for which more significant cooperative efforts, beyond discussions in dialogues, are both important and tractable in the current context of US–China competition.

Readers diving further into our report can skip to the following sections based on their interest:

- **Section 2** provides brief background information about the broader US-China AI relationship.
- **Section 3** presents the results of our analysis of over 40 key primary AI policy and corporate governance documents from the US and China, across a range of different risks from AI (e.g., limited user transparency, poor reliability) and governance approaches (e.g., setting up new institutions), which are graded by the extent of overlap between American and Chinese perspectives.
- **Section 4** presents our recommendations in detail, based on analysis and information from the prior two sections

Contents

List of Acronyms

General

- UN Convention on the Law of the Sea (UNCLOS)
- Exclusive Economic Zone (EEZ)
- International Dialogues on AI Safety (IDAIS)
- The United Nations General Assembly (UNGA)
- Intellectual Property (IP)
- Chemical, Biological, Radiological or Nuclear (CBRN)
- Floating Point Operations (FLOPs)
- Floating Point Operations per Second (FLOP/s)
- Infrastructure as a Service (IaaS)
- Coalition for Content Provenance and Authenticity (C2PA)
- Privacy-Enhancing Technologies (PETs)

US

- Federal Trade Commission (FTC)
- Center for Security and Emerging Technology (CSET)
- National Security Agency (NSA)

- National Institute of Standards and Technology (NIST)
  - NIST's AI Risk Management Framework (RMF)
- Department of Homeland Security (DHS)
- Department of Energy (DOE)
- Department of Health and Human Services (DHHS)
- Defense Production Act (DPA)
- Cybersecurity and Infrastructure Security Agency (CISA)
- Executive Order (EO)
  - Biden administration's 2023 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (AI EO)

China

- Standardization Administration of China (SAC)
- Chinese Communist Party (CCP)
- National Information Security Standardization Technical Committee (TC260)
- China Academy of Information and Communications Technology (CAICT)
- The Ministry of Industry and Information Technology (MIIT)
- The Cyberspace Administration of China (CAC)
- Chinese Academy of Social Science (CASS)
- China University of Political Science and Law (CUPL)
- Center for International Security and Strategy (CISS) of Tsinghua University

1 Introduction

Artificial intelligence (AI) systems developed in one country can profoundly shape outcomes around the world. As such, global cooperation is needed to effectively manage the transnational nature of AI development and deployment, especially in areas such as risk mitigation and safety standards.[4] Cooperation between the United States and China is particularly important as the two countries lead in global AI investment, research, development, and deployment. While strategic competition and disagreements between the two nations risk undermining cooperative efforts, certain shared interests persist.

Leaders in both the US and China have already acknowledged the need for governance to address potential AI risks. The Biden administration's 2023 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Biden AI EO) emphasizes that to manage substantial risks from AI "policies, institutions, and, as appropriate, other mechanisms" are required to test and evaluate systems before they are deployed."[5] Similarly, a resolution adopted at the third plenary session of the 20th Central Committee of the Communist Party of China states the CCP's intention to "improve the mechanisms for developing and managing generative artificial intelligence" and "institute oversight systems to ensure the safety of artificial intelligence."[6]

Both countries have also recognized the importance of international cooperation on AI governance and safety. They each signed the Bletchley Declaration at the 2023 AI Safety Summit, China has also co-sponsored the US's 2024 United Nations General Assembly resolution to promote safe, secure, and trustworthy AI systems for sustainable development. These official initiatives have been complemented by Track II diplomacy efforts, such as the International Dialogues on AI Safety, a series of discussions between scientists from both nations.

However, significant obstacles to US–China cooperation on AI issues remain. In the midst of current geopolitical tensions, leading in the development of advanced AI systems is viewed by both nations as critical for both national security and economic competitiveness. Accordingly, the US and China continue to make substantial investments in domestic AI projects, while seeking to control critical inputs for AI development. For example, the US has implemented a series of export controls to restrict Chinese firms' access to advanced semiconductors crucial for AI development.

Yet even in the presence of fundamental disagreements, states have been able to cooperate on issues of mutual interest in the past.[7] For example, during the Cold War, the US and Soviet Union cooperated on nuclear safety and security issues, so as to prevent unintended

[4] Hadrien Pouget et al., "The Future of International Scientific Assessments of AI's Risks," *Oxford Martin AI Governance Initiative,* August 2024, https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-international-scientific-assessments-of-ais-risks

[5] "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210) (2023): 75191–75226, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[6] Nicholas Welch, "Tech Policy at the Third Plenum," *ChinaTalk*, August 2024.

[7] Seán Ó hÉigeartaigh et al., "Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance," *Centre for the Study of Existential Risk*, May 2020, https://www.cser.ac.uk/resources/cross-cultural-cooperation/

nuclear detonations.[8]

This paper aims to inform cooperation on AI safety and governance between the US and China, by identifying topics suitable for dialogue between the US and China on AI governance and safety.

It is structured as follows:

- **Section 2** provides brief background information about the broader US-China AI relationship. Information from Section 2 is taken into consideration further on in Section 4, informing our understanding of how potential dialogue topics intersect with the existing US-China AI relationship.
- **Section 3** presents the results of our analysis of over 40 key primary AI policy and corporate governance documents from the US and China, across a range of different risks from AI (e.g., limited user transparency, poor reliability) and governance approaches (e.g., setting up new institutions), which are graded by the extent of overlap between American and Chinese perspectives.
- **Section 4** incorporates the results focusing on whether risks and governance approaches that show strong or moderate overlap in the prior section (i.e. areas where there is some shared understanding about AI safety and governance concepts between the US and China), could be suitable topics for US-China AI dialogue. In developing our recommendations, this section sets aside topics where there are no clear ways for international cooperation to take place, and topics which could be deemed too sensitive for cooperation (e.g., linked to existing tensions between the two countries).
- **Section 5** lays out some of the limitations of our analysis, while **Section 6** concludes the paper.

[8] Jeffrey Ding, “Keep your enemies safer: technical cooperation and transferring nuclear safety and security technologies,” European Journal of International Relations, April 2024, https://journals.sagepub.com/doi/10.1177/13540661241246622.

2 Background: The US–China AI relationship

This section provides relevant background about the US-China AI relationship, namely through a review of existing comparative analyses of US and Chinese AI policy, and a brief outline of recent international cooperation on AI safety and governance involving both the US and China.

For readers less familiar with the US-China bilateral relationship, Appendix A provides an overview of non-AI dimensions of the relationship that also informs some parts of our analysis, covering politics and values, geopolitics and security, economics and trade, and climate and society. While US-China relations have become more fraught in recent years, Appendix A shows that there has been consistent, if uneven and halting, progress.

2.1 Existing analyses

In this section, we review recent papers that have taken a comparative approach to studying US and Chinese AI regulation. Chun et al. (2024) compare the US, Chinese and EU regulatory approaches to AI, focusing on major points of divergence rather than convergence.[9] The authors highlight differences in risk-benefit trade-offs, in the optimal balance between safety versus innovation and cooperation versus competition, and in trust in centralized authority versus trust in market-driven regulation. They contrast the market-driven and decentralized US system with the more centralized Chinese approach, which synthesizes use-case specific laws with general guidelines translated into a centralized registration, testing, and monitoring framework.

Giardini and Fritz (2024) analyze the extent to which AI regulations in various countries, including the US and China, align with the OECD AI Principles.[10] Examining convergences, they find that the US and Chinese approaches align on issues such as purported respect for the rule of law and human rights; algorithmic discrimination; cybersecurity requirements; data security and traceability requirements. They also find that neither country has established an AI incident notification system. The analysis also notes some divergences. For example, the US has more detailed requirements for system safety and testing, while in contrast, China's regulations on content moderation are more sweeping.

Hine and Floridi (2022), in a paper that combines philosophical and quantitative document analyses, find some intractable disagreements between the US and China, such as incompatible desires to be a clear world leader in AI.[11] However, they also find some areas

[9] Jon Chun, Christian Schroeder de Witt and Katherine Elkins, "Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US," October 2024, https://arxiv.org/pdf/2410.21279

[10] Tommaso Giardini and Johannes Fritz, "The Anatomy of AI Rules: A Systematic Comparison of AI Rules Across the Globe," *Digital Policy Alert*, June 2024, https://digitalpolicyalert.org/ai-rules/the-anatomy-of-AI-rules; "OECD AI Principles Overview," *Organisation for Economic Co-operation and Development*, https://oecd.ai/en/ai-principles

[11] Emmie Hine and Luciano Floridi, "Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies," January 2022, https://ssrn.com/abstract=4006332

of agreement, including a desire to realize a “Good Domestic AI Society” by sharing the benefits of the technology widely. Hine (2023) applies the same analytical techniques to understand more recent developments.[12] She concludes that the US's unyielding commitment to “democratic values” is a serious impediment to agreement on a global, pluralistic approach to AI governance, though grounding future discussions in human rights treaties with nominal international support could facilitate future cooperation.

Zeng et al. (2024) carry out a categorization task similar to the one utilized in this paper, collating government and corporate documents from the US, EU and China.[13] Their focus, however, is on creating a comprehensive AI risk taxonomy, examining the similarities and differences in how risk is understood in the three contexts with the aim of developing a common language for information sharing and the promotion of best practices in risk mitigation.

2.2 Evidence from diplomatic processes

Turning to direct evidence from multilateral fora, the Bletchley Declaration, signed by both the US and China, along with the 26 other countries attending the 2023 AI Safety Summit and the European Union, indicated some consensus that advanced AI presents both opportunities and a range of significant risks. Specific risks mentioned include threats to human rights, privacy, and fairness, as well as the potential for “catastrophic” harm from intentional misuse or loss of control of advanced AI systems. The Declaration urges a proportionate, pro-innovation response from governments that balances these risks and benefits. This consensus however, was not extended at the Seoul Summit in May 2024, as while China did attend, it did not sign onto the intergovernmental statement.[14] At the same summit, however, Zhipu AI, a leading Chinese AI developer, did sign onto the Seoul Commitments, a set of corporate commitments made by companies. Corporate signatories pledged to test their models for concerning capabilities and implement measures to mitigate risks from these systems, including halting development or deployment if necessary.

The US and China also both supported the General Assembly in adopting the 2024 resolution “Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development”. This wide-ranging document indicates multiple areas of overlap, including the potential for AI systems to address major global issues such as achieving the Sustainable Development Goals and promoting peace; that AI could hinder progress towards achieving various development goals by widening digital divides, reinforcing structural inequalities and biases, and undermining human rights, among other threats; that there is a need for international cooperation to ensure “interoperable” standards and practices; and that discussions on appropriate governance approaches should continue and be accelerated.

[12] Emmie Hine, “Governing Silicon Valley and Shenzhen: Assessing a New Era of Artificial Intelligence Governance in the US and China,” August 2023, https://ssrn.com/abstract=4553087

[13] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia Dawn Song, Percy Liang, and Bo Li, “AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies,” June 2024, https://arxiv.org/pdf/2406.17864

[14] Jessica Birch and Öykü Özfırat, “Key Takeaways from the AI Seoul Summit 2024,” *Access Partnership*, May 2024 https://accesspartnership.com/key-takeaways-from-the-ai-seoul-summit-2024/

There has also been progress on the bilateral front. Following on from the joint statement after the Biden-Xi summit at the end of 2023, China and the US also held a round of intergovernmental talks on AI in Geneva in May 2024, with a second round of talks announced in August 2024.[15] In November 2024, the heads of state of both countries reached an agreement that AI systems should not be given control over the decision to use nuclear weapons.[16]

Several "Track II" dialogues between US and Chinese scientists have sought to extend and clarify these areas of agreement. One set of dialogues was initiated by Henry Kissinger during his visits to China in 2023, where he met with Chinese President Xi Jinping, in part to discuss the need for international cooperation at the highest levels to mitigate serious risks from AI.[17] A separate set of dialogues, The International Dialogues on AI Safety (IDAIS) were initiated by senior scientists and have resulted in two consensus statements. First, the Ditchley Statement, which preceded the AI Safety Summit, stated that "coordinated global action" is "critical" to prevent unacceptable AI risks, highlighting misuse risks and loss of control as particularly concerning. A follow-up dialogue in Beijing in 2024 resulted in a significantly longer statement that mentioned specific red lines AI systems should not be allowed to cross, such as autonomous replication, power seeking, assistance in developing weapons, executing cyberattacks, or deceiving designers or regulators. A third dialogue in Venice in late 2024 resulted in a statement calling for AI safety to be recognized as a global public good, as well as global emergency preparedness agreements and institutions, amongst other proposals. On the whole, the IDAIS statements highlight major risks from AI systems causing harm or escaping human control as particularly promising issues for future dialogues to address.[18]

Considering the breadth and potential severity of AI risks indicated in the Bletchley Declaration and UNGA Resolution, as well as the importance of international cooperation in addressing them, there is a clear need to identify the specific issues on which progress towards meaningful US–China agreement is possible.

[15] Lim Min Zhang, "US, China agree to expand military talks, continue AI cooperation after Sullivan-Wang meet," *The Straits Times,* August 2024, https://www.straitstimes.com/asia/east-asia/us-china-agree-to-expand-military-talks-continue-ai-cooperation-after-sullivan-wang-meet

[16] Jarrett Renshaw and Trevor Hunnicutt, "Biden, Xi agree that humans, not AI, should control nuclear arms," *Reuters,* November 2024, https://www.reuters.com/world/biden-xi-agreed-that-humans-not-ai-should-control-nuclear-weapons-white-house-2024-11-16/

[17] Bob Davis, "Back on Track?," *The Wire China,* January 2024, https://www.thewirechina.com/2024/01/14/back-on-track-two-dialogues-u-s-china-dialogues/.

[18] International Dialogues on AI Safety, https://idais.ai/.

3 Results

In this section we share some of the results drawn from our analysis of domestic policy and corporate governance documents. We analyzed 44 documents issued by actors such as the US government, US AI developers, the Chinese government, Chinese legal experts, and Chinese AI developers. We read through these documents, extracting relevant quotes and coding each quote using a taxonomy adapted from CSET's AGORA project. We focus specifically on analysing how these documents discuss risks from AI and governance approaches. For each risk or governance approach, we compared the quotes coded in both the US and Chinese sources, assessing whether concepts discussed were similarly emphasized or discussed. As much as possible, we attempted to add nuance or context based on the broader provenance of a quote (e.g., where it was placed in a text, or what larger document it was part of), instead of just relying on a specific extract or quote alone. For each risk and governance approach to identify areas of strong, moderate, and weak overlap, defined below.

Table 1: Criteria for assessing degrees of overlap between US and Chinese documents

Degree of overlap	**Criteria**
Strong	Most, if not all, of the material found in the dataset shows concurrence between US and Chinese understanding of a given risk or governance approach
Moderate	While there is some overlap in understanding, there are also key differences that we could identify in the way concepts were raised
Weak	Majority of the evidence pointed towards differing understandings of emphases on a given issue

A full description of the methodology describing our processes for data collection, coding, and analysis is available in Appendix B.

It is worth noting that sources within a country do in some cases disagree with each other or provide different perspectives. Where possible we document the different conceptions of different risks and governance approaches mentioned by different actors. We do not attempt to comprehensively characterise intra-country differences in perspectives (e.g., whether corporate and government actors tend to have different views), but believe it could be a worthwhile direction for future work.

3.1 Summary

With respect to risks, our analysis suggests that there is at least moderate overlap on all major risk categories we tracked, except privacy, which we classify as having 'weak' overlap.

We observe:

- Strong overlap with risks of limited user transparency and poor reliability.
- Moderate overlap in risks from lack of robustness, bias and discrimination, lack of interpretability and explainability, dangerous capabilities (i.e. critical, large-scale and national security-relevant capabilities), as well as weak cybersecurity.
- Weak overlap in privacy risks.[19]

Table 2: Risk Definitions

Risk	**Definition**
Limited user transparency	Risks related to scenarios where individuals or entities do not have access to information about inputs into particular decisions and / or critical details about the AI system they are interacting with.
Poor reliability	Risks related to the inability of an AI system to perform as required, under the conditions of expected use and over a given period of time, including the entire lifetime of the system.
Lack of robustness	Risks related to an AI system's inability to function as intended under unexpected or unusual circumstances, such as when encountering adversarial inputs or data outside the training distribution.
Bias and discrimination	Risks related to undesirable biases in the outputs of AI systems, including biases according to commonly protected classes such as race or gender.
Lack of interpretability and explainability	Risks stemming from opacity of mechanisms underlying an AI system's operation and/or the meaning of the systems' output in the context of use.
Dangerous capabilities	Risks related to AI system capabilities that could pose critical, large-scale, and national security-relevant threats, including CBRN, cyber, persuasion, autonomy and self-replication.
Weak cybersecurity	Risks related to the security of digital systems associated with AI, including systems involved in development and training, systems housing related intellectual property, and computing infrastructure for deployed models.
Lack of privacy	Risks related to the unauthorized use, disclosure or sharing of personally identifying information in an AI system's inputs or outputs.

[19] Our analysis of attitudes towards privacy is likely limited as we did not analyze existing non-AI specific discussions on privacy, or regulations such as the Personal Data Protection Law.

With respect to governance approaches, we find that there is less overlap, with the majority of the governance approaches showing moderate or weak overlap.

We observe:

- Strong overlap in the recognition of the pro-safety role that AI systems can play and the need for convening stakeholders from different backgrounds.
- Moderate overlap in two broad categories:
  - Governance – creating specific governance mechanisms for AI, requiring external auditing as well as licensing and registration of some form
  - Technical safety – developing technical technical solutions such as watermarking, model evaluations, adversarial testing
- Weak overlap in other governance approaches ranging from risks assessments and disclosure of key dimensions of model information, to input controls, liability, as well as pilots and testbeds.

Table 3: Governance Approach Definitions

Governance approach	Definition
Use of AI for pro-safety purposes	Involving the use of AI systems to improve AI safety (e.g., using AIs for detection of model anomalies)
Convening	Facilitating, requiring, setting conditions on, or otherwise addressing the convening of different stakeholders to tackle governance challenges - for example, to share feedback or to participate in governance development.
Governance development	Creating, supporting, or requiring the development of public or private governance mechanisms or institutions related to the development or deployment of AI systems.
Technical solutions	Developing technical solutions as a strategy to govern particular risks (e.g., content labelling and watermarking)
Evaluations	Requiring, or encouraging, the systematic evaluation of AI systems
Adversarial testing	Evaluation in which the evaluator takes an adversarial approach, seeking to subvert or otherwise produce undesirable results from the system or process being tested by any means available. Sometimes called ”red teaming.”
External auditing	Evaluation by a disinterested counterparty or third party, such as a customer or a professional auditing firm.
Licensing or registration	Requiring, incentivizing, or otherwise encouraging actors involved in AI-related activities, such as AI developers, vendors, users, or researchers, to either receive sanction from a regulator for their activities (licensing) or to notify a regulator of their activity pursuant to a formal process (registration).
Risk assessments	Identification, assessment and in some cases quantification of potential sources of and pathways to harm caused by AI systems.
Disclosure	Requiring or encouraging, the disclosure of information about AI systems by their users, developers, vendors, or other stakeholders to third parties, including but not limited to the general public.
Input controls	Restricting or placing conditions on the sale, distribution, or use of technical inputs to AI systems, specifically data or computational resources.
Liability	Holding specific parties accountable through civil or criminal penalties for unlawful actions
Pilot and testbeds	Creating, facilitating, setting conditions on, or otherwise addressing the development and operation of government-supported or government-conducted pilot programs or test environments related to artificial intelligence.

3.2 Risks

3.2.1 Strong overlap

Limited user transparency: Risks related to scenarios where individuals or entities do not have access to information about inputs into particular decisions and / or critical details about the AI system they are interacting with.

- The NIST Risk Management Framework from the US describes transparency as foundational for accountability, using it to refer to the extent to which information about an AI system and its outputs is available to a user.[20]
- Chinese documents largely have a similar understanding. An AI safety standard (TC260-003) calls for more transparency, with specific requirements for disclosure of certain contextual information about models.[21]Older Chinese recommendation algorithm regulations from 2021 refer to transparency in a slightly different context, with requirements on service providers to provide users with information about rules used for content recommendation, similar to the NIST definition above.[22]

Poor reliability: Risks related to the inability of an AI system to perform as required, under the conditions of expected use and over a given period of time, including the entire lifetime of the system.

- NIST, via its Risk Management Framework in the US, frames reliability as 'the ability of an item to perform as required, without failure', calling reliability a goal for overall correctness of an AI system.[23]
- The Ministry of Industry and Information in China and a related think-tank (the CAICT) both mention reliability in a similar way, and further elaborate on it as an

[20] "Trustworthy AI depends upon accountability. Accountability presupposes transparency. Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system. . . " "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology U.S. Department of Commerce, January 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[21] "Service transparency: ... If the service is provided using an interactive interface, the following information shall be disclosed to the users...: Limitations of the service; Summary information on the models and algorithms used, etc.; The personal information collected and its uses in the service." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China, February 2024, https://perma.cc/GU3Q-GAJ3 (Translated by the Center for Security and Emerging Technology, April 2024, https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/).

[22] "The providers of algorithmic recommendation services are encouraged to comprehensively use strategies such as for eliminating duplicate content and for fragmentation and intervention, and optimize the transparency and explainability of rules for searches, sorting, selections, pushing, and displays, to avoid producing a negative impact on users, and to prevent and reduce contention and disputes." "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," Cyberspace Administration of China, December 2021, http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm (Translated by China Law Translate, January 2022, https://www.chinalawtranslate.com/en/algorithms/)

[23] "Reliability is defined in the same standard as the 'ability of an item to perform as required, without failure, for a given time interval, under given conditions'. . . Reliability is a goal for overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system." "AI RMF 1.0," National Institute of Standards and Technology U.S. Department of Commerce.

area to be standardized, with specific reliability testing required under a machine learning cybersecurity standard.[24]

3.2.2 Moderate overlap

Lack of robustness: Risks related to an AI system's inability to function as intended under unexpected or unusual circumstances, such as when encountering adversarial inputs or data outside the training distribution.

- In the US, NIST's Risk Management Framework refers to robustness as 'the ability of a system to maintain its level of performance under a variety of circumstances'.[25]
- Some understanding of robustness is apparent in Chinese sources, though it is unclear if the exact same definition is in use. Alibaba's 2023 AI Governance whitepaper and a 2021 industry whitepaper both refer to robustness as part of new issues brought about by the black box nature of generative AI models.[26]

Bias and discrimination: Risks related to undesirable biases in the outputs of AI systems, including biases according to commonly protected classes such as race or gender.

- Bias and discrimination is brought up by both Western and Chinese sources in some similar ways, but with notable differences.
- The White House & FTC in the US, as well as the Chinese standards-setting body TC260, Baichuan and Alibaba have all variously expressed concerns around discrimination and disinformation leading to societal harm.[27]

[24] "...Standardize the technical research and development and operational service requirements of artificial intelligence, including technical requirements and evaluation methods for artificial intelligence robustness, reliability, traceability, artificial intelligence governance support technology...The ethical governance requirements of the cycle include artificial intelligence ethical risk assessment, artificial intelligence fairness, explainability and other ethical governance technical requirements and evaluation methods, artificial intelligence ethical review and other standards." "Guidelines for the Construction of a National Comprehensive Standardization System for the Artificial Intelligence Industry," Ministry of Industry and Information Technology, January 2024, https://www.miit.gov.cn/gzcy/yjzj/art/2024/art_983199be076649d494690135c0b4d168.html.

[25] "Robustness or generalizability is defined as the 'ability of a system to maintain its level of performance under a variety of circumstances'... Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting." "AI RMF 1.0," National Institute of Standards and Technology U.S. Department of Commerce.

[26] "Model security refers to the inherent security of generative model, which mainly includes two aspects: traditional software and information technology security issues, such as backdoor vulnerabilities, data theft, reverse engineering on the one hand; new security issues brought about by the black box model characteristics of the artificial intelligence technology, such as fairness, robustness, explainability, on the other hand." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence," October 2023.https://mp.weixin.qq.com/mp/appmsgalbum?__biz=Mzg4MTY2MzUyNA==&action=getalbum&album_id=3187743423251611652&scene=173&from_msgid=2247572341&from_itemidx=1&count=3&nolastread=1#wechat_redirect, "White Paper on AI Security Evaluation," National Quality Supervision and Inspection Center for Speech and Image Recognition Products.

[27] "At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disin-

- These documents share a similar scope, with both discussing discrimination/bias along some similar lines, notably race/ethnicity, sex, and religion/beliefs.[28]

- US documents, particularly many parts of the Biden AI EO, discuss bias in relation to model use in decision-making in practice - including the responsibility of the decision-maker, the vendor, etc to ensure adequate measures are in place to mitigate bias.[29]

- Chinese companies similarly point to the risk that bias introduces in decision making processes.[30]

- However, bias and discrimination is contextual and therefore understood in different ways by different actors.

- For example, while AI labs from both China and the West test their models against bias benchmarks, labs like Baichuan acknowledge that their models may not fully account for biases relevant in non-Chinese contexts.[31]

formation; displace and disempower workers; stifle competition; and pose risks to national security." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210);"Evidence already exists that fraudsters can use [AI] tools to generate realistic but fake content quickly and cheaply, disseminating it to large groups or targeting certain communities or specific individuals..." Staff in the Office of Technology and The Division of Privacy and Identity Protection, "AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive," Federal Trade Commission, February 2024, https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive;"The personal characteristics derived from feature generation include content such as obscenity, pornography, gambling, superstition, terror, violence, etc., or content expressing discrimination against ethnicity, race, religion, disability, disease, etc." "Standards Regarding Security Requirements for Automated Decision-Making Based on Personal Information," National Technical Committee 260 on Cybersecurity of Standardization Administration of Chin; "There's also the potential for misuse, as the model could be used to generate harmful or misleading content. Although we try our best efforts to balance safety and utility, some safety measures may appear as over-cautions, affecting the model's usability for certain tasks." Baichuan Inc., "Baichuan 2: Open Large-scale Language Models," September 2023, https://cdn.baichuan-ai.com/paper/Baichuan2-technical-report.pdf; "[Transcription of figure] Breaks down the idea of AI security into four categories. The first is personal information, the second is model security which includes bias, wrong value, adversarial attack, and induced attack. The third is content security referring to porn, gambling, drug abuse, general abuse, and illegal advertising. The final category is intellectual property, which includes copyright and the right to use as well as legal protections afforded to AI-generated content." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence."

[28] "If use of an algorithmic decision-making tool has an adverse impact on individuals of a particular race, color, religion, sex, or national origin, or on individuals with a particular combination of such characteristics..., then use of the tool will violate Title VII unless the employer can show that such use is 'job related and consistent with business necessity' pursuant to Title VII." "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," U.S. Equal Employment Opportunity Commission, Title VII, 29 CFR Part 1607 (2023)https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial; "During processes such as algorithm design, the selection of training data, model generation and optimization, and the provision of services, effective measures are to be employed to prevent the creation of discrimination such as by race, ethnicity, faith, nationality, region, sex, age, profession, or health," "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China, July 2023, http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[29] ". . . a meeting of the heads of Federal civil rights offices...to discuss comprehensive use of their respective authorities and offices to: prevent and address discrimination in the use of automated systems, including algorithmic discrimination. . . " "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[30] "At the application level, the risk involves algorithmic bias, ethical conflict, and labor substitution among others. For example, due to subjective factors, or bias contained in training data and data input during self-learning process, machine-learning algorithms may introduce bias into its decision-making process." "AI Governance Whitepaper," SenseTime https://perma.cc/SNU4-VKVU

[31] "Like other large language models, Baichuan 2 also faces ethical challenges. It's prone to biases and toxicity, especially given that much of its training data originates from the internet... While optimized for Chinese and English for safety, the model has limitations in other languages and may not fully capture biases relevant to non-Chinese cultures." Baichuan Inc., "Baichuan 2: Open Large-scale Language Models.";"To understand bias and stereotyping in text-to-text capabilities, we focus on the Winogender (Rudinger et al., 2018), Winobias (Zhao et al., 2018), and Bias Benchmark in QA (BBQ) (Parrish et al., 2021);" Gemini Team Google, "Gemini: A Family of Highly Capable Multimodal Models,"

Lack of interpretability and explainability: Risks stemming from opacity of mechanisms underlying an AI system's operation and/or the meaning of the systems' output in the context of use.

- US and Chinese documents both acknowledge the importance of ensuring interpretability and explainability (I&E).
- One key difference is that the US appears more optimistic about interpretability and explainability–linked governance mechanisms, while Chinese documents focus on I&E as an inherent problem for AI systems.
- However, this does not mean that Chinese entities view I&E as unsolvable.
- In the US, documents from NIST mention interpretability and explainability.[32]
- Meanwhile, several Chinese documents, from various actors (Alibaba, CAC, SAC, MIIT) bring up interpretability and explainability.[33]
- Chinese actors tend to use the terms to describe the problems of the black box nature of algorithms and inherent challenges for AI safety, with Sensetime even referring explicitly to an "interpretability risk" from our limited ability to understand algorithms.[34]
- That said, MIIT does link interpretability and explainability to governance mechanisms.

December 2023, https://arxiv.org/abs/2312.11805; "We continue to make good progress on improving our models' performance in situations that could lead to bias and discrimination. On our recently released evaluation for discrimination [72], Claude 3 Opus and Sonnet exhibit comparable discrimination scores to Claude 2.1, and Claude 3 Haiku has comparable or lower scores compared to Claude Instant 1.2. The discrimination score indicates how different (in logit space) the models' likelihood of recommending a positive decision is to different subjects across 10 different demographic characteristics spanning race, gender, and age." Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," March 2024, https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf.

[32] "Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes." "AI RMF 1.0," National Institute of Standards and Technology U.S. Department of Commerce.

[33] "Model security refers to the inherent security of generative models, which mainly includes two aspects: traditional software and information technology security issues, such as backdoor vulnerabilities, data theft, reverse engineering on the one hand; new security issues brought about by the black box model characteristics of artificial intelligence technology, such as fairness, robustness, explainability, on the other hand." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence"; "The providers of algorithmic recommendation services are encouraged to... optimize the transparency and explainability of rules for searches, sorting, selections, pushing, and displays...." "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," Cyberspace Administration of China; "The algorithm logic or model behind the feature generation of computer programs cannot be explained or cannot be explained clearly;" "Standards Regarding Security Requirements for Automated Decision-Making Based on Personal Information," National Technical Committee 260 on Cybersecurity of Standardization Administration of China; "... standardize the technical research and development and operational service requirements of artificial intelligence, including technical requirements and evaluation methods for artificial intelligence robustness, reliability, traceability, artificial intelligence governance support technology; ... The ethical governance requirements of the cycle include artificial intelligence ethical risk assessment, artificial intelligence fairness, explainability and other ethical governance technical requirements and evaluation methods, artificial intelligence ethical review and other standards." "Guidelines for the Construction of a National Comprehensive Standardization System for the Artificial Intelligence Industry," Ministry of Industry and Information Technology.

[34] "At the algorithm level, the risk mainly involves algorithm decision-making, black box algorithm, and algorithm security. Of these, algorithm decision-making risk refers to the inability to predict the reasons and effects of decisions made by AI systems, due to the unpredictability of the results of algorithmic reasoning and the cognitive limitations of human beings. For example, the problem of liability fixation is a typical one. Black box algorithm risk refers primarily to interpretability risk resulting from opaque decision-making and the inability to be fully explained due to the complexity of neural network algorithms..." "AI Governance Whitepaper," SenseTime.

Dangerous capabilities: Risks related to AI system capabilities that could pose critical, large-scale, and national security-relevant threats, including CBRN, cyber, persuasion, autonomy and self-replication.

- Western labs clearly state the need for dangerous capability evaluations in their technical reports, model cards, and responsible scaling policies.[35]
- Chinese labs have shown awareness of these developments in whitepapers.[36]
- Chinese state-linked think-tanks running model evaluations have begun to include some types of dangerous capability evaluations as part of their benchmarks.[37]
- With regard to dangerous capabilities linked to chemical, biological, radiological or nuclear (CBRN) weapons development, there is quite strong overlap.
  - The US very often mentions CBRN risks, focusing on biological and chemical domains.[38]
  - Chinese mention of bio-chemical risks also exist albeit not to the same degree as in the US. An updated AI safety standard by the TC260 cybersecurity standardization committee included mention of the use of AI to create chemical or biological weapons in 2024 (though this was removed in a later draft).[39]
  - Beyond that, hazardous chemicals are one of 20+ things CAICT is evaluating models for.[40]
  - It does appear that Chinese actors tend to still see this under the lens of 'dangerous content' as the CAICT report puts hazardous chemicals under the category content security and subcategory violation of laws and regulations, alongside things like gambling and pornography.

[35] "Adversarial Testing via Domain Experts. . . To understand the extent of these risks, we engaged over 50 experts from domains such as long-term AI alignment risks, cybersecurity, biorisk, and international security to adversarially test the model." OpenAI, "GPT-4 Technical Report," March 2023, https://arxiv.org/abs/2303.08774.

[36] "As the capabilities of large AI models continue to grow, it is crucial to assess potential risks in the entire life cycle of AI models. Many experts believe that AI will eventually be able to perform most human tasks, including technology development and business operations. But there are concerns that AI systems could become uncoordinated and pursue goals harmful to civilization, leading to potential risks of global catastrophe. Therefore, model evaluation becomes increasingly important. Leading AI companies such as OpenAI and Anthropic are evaluating their AI systems by themselves or in cooperation with external third-party organizations." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[37] Jeffrey Ding, "ChinAI #261: First results from CAICT's AI Safety Benchmark," ChinAI Newsletter, April 2024.https://chinai.substack.com/p/chinai-261-first-results-from-caicts.

[38] "At a minimum, the Secretary shall develop tools to evaluate AI capabilities to generate outputs that may represent nuclear, nonproliferation, biological, chemical, critical infrastructure, and energy-security threats or hazards." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[39] "Service providers should pay close attention to the long-term risks that generative artificial intelligence may bring, be cautious about artificial intelligence that may have the ability to deceive humans, self-replicate, and self-transform, and focus on the possibility that generative artificial intelligence may be used to write malware, create security risks such as biological or chemical weapons." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[40] Jeffrey Ding, "ChinAI #261: First results from CAICT's AI Safety Benchmark."

  - In contrast, the Biden AI EO frames chemical and biological risks more broadly, linked to the differential risk added by AI systems in aiding actors in developing CBRN weapons, relative to the internet and weighed against the defensive advantages granted by the same AI systems.[41]

- With respect to cyber capabilities of AI systems, there is also some common ground.
  - The Biden AI EO calls attention to cybersecurity as an area that AI could cause harm in, including through the development of autonomous cyber capabilities.[42]
  - In China, the TC260 Basic Safety Requirements for Generative Artificial Intelligence Services refers to the possibility that AI may be used to write malware.[43]
  - However, the US specifically expresses greater concern about inadvertent advancement of adversary cyber capabilities. In an Executive Order prohibiting investments in certain national security technologies and products in specific countries of concern, it expresses concerns about transactions that advance the cyber-enabled capabilities of 'countries of concern', including AI products.[44]

- With regards to persuasion as a dangerous capability, there appears to be limited recognition in China, while there is much more widespread concern in the US.
  - In the US, this is a concern explicitly flagged in the OpenAI Preparedness Framework and other technical documents from AI labs (although not in the Biden AI EO).[45]
  - Despite the dissimilarity, Chinese AI regulations have always been concerned with algorithms that have public mobilization or social opinion properties, suggesting

[41] "...Evaluate the potential for AI to be misused to enable the development or production of CBRN threats, while also considering the benefits and application of AI to counter these threats, including, as appropriate, the results of work conducted under section 8(b) of this order." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[42] "...launching an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities through which AI could cause harm, such as in the areas of cybersecurity and biosecurity." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[43] "2) Model input content shall be continuously monitored for malicious input attacks, such as distributed denial-of-service (DDoS), cross-site scripting (XSS), and injection attacks; 3) Regular security audits shall be conducted on the development framework, code, etc. used, focusing on issues related to open-source framework security and vulnerabilities..." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[44] "The regulations issued under this section shall identify categories of prohibited transactions that involve covered national security technologies and products that the Secretary... determines pose a particularly acute national security threat because of their potential to significantly advance the military, intelligence, surveillance, or cyber-enabled capabilities of countries of concern." "Executive Order 14105: Addressing United States Investments in Certain National Security Technologies and Products in Countries of Concern," Federal Register 88 (154) (2023): 54867-54872.

[45] "Based on our general capability evaluations, we expect GPT-4 to be better than GPT-3 at producing realistic, targeted content. As such, there is risk of GPT-4 being used for generating content that is intended to mislead." OpenAI, "GPT-4 Technical Report.";"Persuasion & deception: We tested whether Gemini Pro and Ultra models could persuade or deceive humans in 1-on-1 dialogue settings in studies with human participants. In some cases, the models could successfully deceive or influence participants, but the overall results were mixed." Gemini Team Google, "Gemini: A Family of Highly Capable Multimodal Models."

that mass persuasion via AI is or could become a concern of the Chinese state as well.[46]

- The US does appear to be concerned about the use of federal data to enhance dangerous capabilities, while an equivalent concern was not found amongst our Chinese sources.
  - The Biden AI EO expresses specific concern that federal data may aid in the development of CBRN or autonomous cyber capabilities, and directs the Chief Data Officer Council to draw up guidelines on performing security reviews.[47] The EO also directs the Department of Energy to develop tools to understand and mitigate security risks from AI.[48]
- With regards to other dangerous capabilities, we find some evidence of Chinese concern around model autonomy, self-replication, and evasion of human control or oversight but such capabilities are undeniably a much more widely discussed issue in the US.[49]

Weak cybersecurity: Risks related to the security of digital systems associated with AI, including systems involved in development and training, systems housing related intellectual property, and computing infrastructure for deployed models.

- Both countries express concerns about weak cybersecurity, which can be roughly categorised into concerns related to theft of model weights, and general concerns about software vulnerabilities.
- There is a common ground on the importance of preventing theft or undesired access to model weights.

[46] "Article 17: Those who provide generative artificial intelligence services with public opinion attributes or social mobilization capabilities shall conduct security assessments in accordance with relevant national regulations and perform algorithm registration and change and cancellation registration procedures in accordance with the 'Internet Information Service Algorithm Recommendation Management Regulations'." "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China.

[47] "It also requires addressing AI systems' most pressing security risks — including with respect to biotechnology, cybersecurity, critical infrastructure, and other national security dangers — while navigating AI's opacity and complexity." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[48] "...develop and, to the extent permitted by law and available appropriations, implement a plan for developing the Department of Energy's AI model evaluation tools and AI testbeds." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[49] "The term 'dual-use foundation model' means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters..." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210)."Novel capabilities often emerge in more powerful models. Some that are particularly concerning are the ability to create and act on long-term plans, to accrue power and resources ('powerseeking'), and to exhibit behavior that is increasingly 'agentic.'" OpenAI, "GPT-4 Technical Report"; "Service providers should pay close attention to the long-term risks that generative artificial intelligence may bring, be cautious about artificial intelligence that may have the ability to deceive humans, self-replicate, and self-transform, and focus on the possibility that generative artificial intelligence may be used to write malware, create security risks such as biological or chemical weapons."“Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

  - The Biden AI EO and White House voluntary commitments, in combination, oblige developers to disclose the ownership and possession of any dual-use foundation model weights, alongside the physical and cybersecurity measures taken to safeguard these from theft.[50]

  - As early as 2021, an industry report from several Chinese AI labs noted the possibility of model theft,[51] with a more recent Tencent Large Model Safety and Security report describing the need to prevent model information leakage within an organization, alongside a broader suite of measures to protect model weights.[52]

  - A Chinese standards-setting body also calls on AI service providers to separate inference and training environments to prevent improper access and information leakage.[53]

- There is less clear common ground on general cybersecurity of systems related to AI development and deployment.

  - The only US documents to mention this topic were the Biden AI EO, which calls on developers to share software vulnerabilities that have been discovered and any known exploits associated with the vulnerabilities, and the Cybersecurity and Infrastructure Security Agency (CISA) notice on the obligation of AI systems to be secure by design, which calls for AI products to be built in a way that reasonably protects against malicious cyberattacks.[54]

[50] "Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding the following: any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats; the ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210);"The companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights." The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI" (2023), https://perma.cc/Q8QS-3AGS.

[51] "Due to the openness of real-life application scenarios, artificial intelligence systems face malicious attacks such as adversarial samples and model theft, and the security and trustworthiness of artificial intelligence have been questioned." "White Paper on AI Security Evaluation," National Quality Supervision and Inspection Center for Speech and Image Recognition Products, National Industrial Information Security Development Research Center, Institute of Artificial Intelligence, October 2021.https://www.realai.ai/ai-research

[52] "Internal model information leakage:... Privatize and control all sensitive code repositories related to large models, to ensure that no one outside the project team can access large model-related code, such as training code, etc. On the other hand, it is also necessary to set up specific administrators and management groups for sensitive code repositories to ensure that there is no horizontal unauthorized access within the project team." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[53] "The training environment and inference environment shall be segregated to avoid data leakage and improper access." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[54] "Prior to the development of guidance on red-team testing standards by NIST pursuant to subsection 4.1(a)(ii) of this section, this description shall include the results of any red-team testing that the company has conducted relating to lowering the barrier to entry for the development, acquisition, and use of biological weapons by non-state actors; the discovery of software vulnerabilities and development of associated exploits; the use of software or tools to influence real or virtual events; the possibility for self-replication or propagation; and associated measures to meet safety objectives;" "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210); "AI software must be Secure by Design [...] Secure by Design "means that technology products are built in a way that reasonably protects against malicious cyber actors successfully gaining access to devices, data, and connected infrastructure." Christine Lai and Jonathan Spring, "Software Must Be Secure by Design, and Artificial Intelligence Is No Exception," Cybersecurity and Infrastructure Security Agency, August 2023, https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception.

  - In China, this issue generally receives more widespread attention. A Tencent Large Model Security and Safety Report calls attention to vulnerabilities in open-source platforms used in parts of the machine learning pipeline, suggesting that that risk component library be constructed to monitor and flag any security risks introduced during training.[55] Government standards documents also flag cybersecurity risks in infrastructure. An AI computing information security platform framework issued by the TC260 standard-setting body argues that platforms must design their own security functions such that the platform itself provides a secure computing environment and does not become a weak link in cyber attacks.[56]

3.2.3 Weak overlap

Lack of privacy: Risks related to the unauthorized use, disclosure or sharing of personally identifying information in an AI system's inputs or outputs.

- While ensuring privacy protections through user consent is a significant concern for government agencies in both the US and China, Chinese agencies seem much more concerned about user protections from *companies.*
- There is overlap between the US and China on protecting personally identifiable information.
  - NIST emphasizes protecting individuals' facets of identity, such as one's body, data, and reputation.[57]
  - The CAC's Deep Synthesis Provisions state that deep synthesis providers must obtain consent for biometric information (eg: faces, voices).[58]

[55] "…Open source components are often exposed to various security issues. For example, well-known software such as PyTorch and Tensorflow have been found to have serious security vulnerabilities several times. Therefore, a machine learning risk component library can be constructed based on various types of security intelligence, and this can be used as one of the access judgment conditions for training tasks. If a component introduced by a training task hits an entry in the risk component library, it will be prohibited from running and an alarm will be pushed to the task leader." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[56] "The AI computing platform's own security functions are designed to provide platform users with a safe computing environment and reduce the risk of the platform becoming a weak link in cyber attacks." "Artificial Intelligence Computing Platform Information Security Framework," National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2023.https://www.tc260.org.cn/front/bzzqyjDetail.html?id=20230515154409898112&norm_id=20221102142806&recode_id=51281.

[57] "Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)." "AI RMF 1.0," National Institute of Standards and Technology U.S. Department of Commerce.

[58] "Deep synthesis service providers and technical supporters shall strengthen the management of training data, employ necessary measures to ensure the security of training data, and where training data includes personal information, they shall comply with relevant provisions on the protection of personal information. Where deep synthesis service providers and technical supports provide functions for editing biometric information such as faces and voices, they shall prompt the users of the deep synthesis service to notify the individuals whose personal information is being edited and obtain their independent consent in accordance with law" "Provisions on the Administration of Deep Synthesis Internet Information Services," Cyberspace Administration of China.

- There is also overlap on need for user consent
  - Standards setting bodies in both the US (NIST) and China (CAC) highlight user consent as a key part of privacy.[59]
- In contrast to the US, where personal data related risks are not as frequently raised, Chinese documents express more concern about abuse of personal data by companies, including calls for:
  - Requiring that companies obtain consent for collecting personal information[60]
  - Making it convenient for users to prevent their information from being used for AI training.[61]
  - Reviewing activity logs of algorithms to ensure that collected personal information has been collected consensually, unless specified by law that the information can be collected.[62]
  - Recommending ethics reviews for AI companies, ensuring that users have a right not to be recommended user-based content.[63]
  - These requirements may be downstream of legislation – the 2021 Personal Information Protection Law – that protects Chinese users from corporate collection of personal data. As we did not look at non-AI specific documents such as this law, our analysis on the similarities and differences with respect to privacy risks may be limited.

[59] "Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)." "AI RMF 1.0," National Institute of Standards and Technology U.S. Department of Commerce; "Where personal information is involved, the consent of the personal information subject shall be obtained or it shall comply with other situations provided by laws and administrative regulations;" "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China.

[60] "Providers shall fulfill confidentiality obligations towards information input by users and users' usage records in accordance with law; they must not collect unnecessary personal information, must not illegally retain user input information and usage records from which users' identities can be determined, and must not illegally provide user input information and usage records to others." "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China.

[61] "When using personal information to provide services such as information push and commercial marketing through machine learning algorithm services, provide algorithm services that do not target personal characteristics or offer individuals easy ways to refuse, and do not induce users through coercion, disguised coercion, frequent reminders, etc." "Information Security Technology – Assessment Specification for Security of Machine Learning Algorithms," Standardization Administration of China; "Users shall be provided with a way to turn off the use of their entered information for training purposes, e.g., by providing the user with options or voice control commands; the turn-off method shall be convenient, e.g., no more than 4 clicks shall be required for the user to reach the option from the main interface of the service when using the options method." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[62] "Review the activity logs of the algorithm's lifecycle to check if the processed personal information has obtained the consent of the subjects of personal information, except where the law or regulations specify that consent is not required."

[63] "Deep synthesis service providers shall implement primary responsibility for information security, establishing and completing management systems such as for user registration, review of algorithm mechanisms, scientific ethics reviews, review for information publication, data security, personal information protection, prevention of telecommunication network fraud, and emergency response, and shall have safe and controllable technical safeguard measures." "Provisions on the Administration of Deep Synthesis Internet Information Services," Cyberspace Administration of China.

3.3 Governance approaches

3.3.1 Strong overlap

Use of AI for pro-safety purposes: Involving the use of AI systems to improve AI safety (e.g., using AIs for detection of model anomalies)

- There is significant overlap in understanding that AI systems can be helpful for safety in a variety of ways.
- In the US, the Biden AI EO calls on the Secretary of Defense and Secretary of Homeland Security to identify and pilot ways to use AI to discover and fix critical vulnerabilities in USG software systems and networks.[64] In calling on the Department of Homeland Security (DHS) to evaluate CBRN threats from AI, the Biden AI EO also asks them to consider the benefits of application of AI to counter CBRN threats.[65]
- On the other hand, in China, a recent whitepaper by Tencent in 2023 identifies that AI can be used to identify and predict cyber threats (e.g., by monitoring web traffic and flagging anomalous patterns).[66] Zhipu, a leading AI startup, in a technical release paper for their GLM-130B model, points to promoting LLM-inclusivity as a way to defend against potential harms, instead of restricting access to LLMs.[67] The CAICT Safety Model Evaluation standard released in 2024 also explicitly points to their use of AI models to assist in running evaluations.[68]

[64] "...Develop plans for, conduct, and complete an operational pilot project to identify, develop, test, evaluate, and deploy AI capabilities, such as large-language models, to aid in the discovery and remediation of vulnerabilities in critical United States Government software, systems, and networks." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[65] "...evaluate the potential for AI to be misused to enable the development or production of CBRN threats, while also considering the benefits and application of AI to counter these threats, including, as appropriate, the results of work conducted under section 8(b) of this order." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[66] "Large models can be used to identify and predict cyber threats, such as malware and cyber attacks. One specific example is the use of large models to analyze network traffic to identify malicious activity. By training on large amounts of network traffic data, large models can learn the patterns and characteristics of various malicious behaviors, such as DDoS attacks, SQL injection, and malware propagation. In this way, security teams can use large models to monitor network traffic and promptly detect and block potential threats." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[67] "While some people think that restricting the access of LLMs can prevent such harmful applications, we argue that promoting LLM inclusivity can lead to better defense against potential harms caused by LLMs. Currently, only governments and large corporations can afford the considerable costs of pre-training LLMs. There is no guarantee that organizations having [sic] substantial financial resources will not do harm using a LLM. Without access to such LLMs, individuals cannot even realize the role of LLMs in the harm." Aohan Zeng et al., GLM-130B: An Open Bilingual pre-Trained Model," Zhipu AI, October 2023, https://arxiv.org/pdf/2210.02414.

[68] "The AI Safety Benchmark evaluation system will cover all fine-grained safety types. After randomly selecting evaluation samples, it conducts automatic assessments using local large safety models." "Authoritative large model AI Safety Benchmark first round results officially released," CAICT, April 2024, https://mp.weixin.qq.com/s/3FcLBHCy_oVaaj-2Ca9zag(Translated by Jeffrey Ding, April 2024, https://docs.google.com/document/d/1WpcydZ6Wv1EwTiS2d9L6j-S5thaeQuziRrOQ0HAiCdk/edit#heading=h.ruzd8o972ahb)

Convening: Facilitating, requiring, setting conditions on, or otherwise addressing the convening of different stakeholders to tackle governance challenges - for example, to share feedback or to participate in governance development.

- There is strong overlap between the US and China when it comes to the convening of different stakeholders to share feedback or participate in governance development. In the US, directives for convening come largely from the Biden AI EO, focus on the following types of convenings in particular: interagency councils, public participation and community engagement, external stakeholder engagement (private sector, academia, civil society, etc.), consulting with experts, labs, AI evaluators, etc..[69]
- In China, regulations such as the Deep Synthesis regulations, call for industry-led self-governance. This convening and collaboration is further evidenced by the list of authors for various Chinese standards, spanning academia, industry and government do suggest that multi-stakeholder convening does take place.[70] Calls for the convening of AI stakeholders appear in draft laws written by experts and legal scholars, rather than codified regulations. For example, the draft law from the Chinese University of Political Science and Law stipulates the creation of an expert committee (including experts in technology, law, and ethics) that provides support for AI safety work.[71] The CASS draft model law also focuses on AI governance mechanisms involving stakeholders from the government, public, and corporate sector.[72]

3.3.2 Moderate overlap

Governance development: Creating, supporting, or requiring the development of public or private governance mechanisms or institutions related to the development or deployment of AI systems.

[69] "Providing Guidance for AI Management. (a) . . . the Director of OMB shall convene and chair an interagency council to coordinate the development and use of AI in agencies' programs and operations, other than the use of AI in national security systems." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210); "To help ensure that people with disabilities benefit from AI's promise while being protected from its risks, including unequal treatment from the use of biometric data like gaze direction, eye tracking, gait analysis, and hand motions, the Architectural and Transportation Barriers Compliance Board is encouraged, as it deems appropriate, to solicit public participation and conduct community engagement;" "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210); "solicit input from the private sector, academia, civil society, and other stakeholders through a public consultation process on potential risks, benefits, other implications, and appropriate policy and regulatory approaches related to dual-use foundation models for which the model weights are widely available." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[70] Relevant industry organizations are encouraged to strengthen industry discipline, establishing and completing industry standards, norms, and systems for self-discipline and management, urging and guiding deep synthesis service providers and technical supports to improve operational specifications, carry out operations in accordance with law, and accept societal oversight." Provisions on the Administration of Deep Synthesis Internet Information Services," Cyberspace Administration of China.

[71] "The main oversight departments for AI organize and establish an expert committee on AI consisting of experts in technology, law, ethics, social [science], and medicine, and other experts, to provide consultation, assessment, validation, and other professional support for AI safety and security work." "Scholars' Draft Law on AI," *China Law Society.*

[72] The State shall establish and improve an artificial intelligence governance mechanism involving government administration, corporate responsibility, industry self-regulation, societal oversight, and user self-discipline, promoting collaborative governance among multiple stakeholders." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences, April 2024, https://doi.org/10.5281/zenodo.10974163.

- In general, it is apparent that there is a lot of energy behind developing governance systems for AI in both China and the US. Here we focus on approaches to international governance development, domestic governance development and the creation of new institutions.

- Both the US and Chinese governments have stated that they want to collaborate with international partners in governing AI, but given the broad scope of such statements, it seems unclear if there is any actual overlap here. For instance, the Biden AI EO states that the US will seek partnerships with other countries on building safeguards,[73] while the Cyberspace Administration of China (CAC) has said that China will "carry out international exchanges and cooperation in an equal and mutually beneficial way" with other nations.[74]

- Both the US and China also call for the development of new governance approaches domestically too. The Biden AI EO calls upon parts of the government to develop new frameworks and institutions,[75] while Chinese documents, such as the two AI law proposals, do also call for the establishment of various governance mechanisms and in the case of the Chinese Academy of Social Science (CASS) draft model law, call for the establishment of a new centralized administration for managing AI governance.[76]

- Differences here include the fact that some of the Chinese documents suggest that there is more obvious pressure put on companies to be responsible to their users. This focus on the responsibility of companies to users is largely absent from the US texts on governance development.

- The US currently appears to be much more focused on creating new institutions, mostly within the executive branch. Besides the US AI Safety Institute within the Department of Commerce, this also includes the establishment of the four new National AI Research Institutes, a coordinating office within the Department of Energy, a task force at the Department of Health and Human Services (DHHS), and a Research Coordination Network focused on advancing privacy research.[77]

[73] "This leadership is not measured solely by the technological advancements our country makes. Effective leadership also means pioneering those systems and safeguards needed to deploy technology responsibly — and building and promoting those safeguards with the rest of the world. My Administration will engage with international allies and partners in developing a framework to manage AI's risks, unlock AI's potential for good, and promote common approaches to shared challenges." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[74] "[Article 6: Encourage independent innovation in basic technologies for generative AI such as algorithms, frameworks, chips, and supporting software platforms, carry out international exchanges and cooperation in an equal and mutually beneficial way, and participate in the formulation of international rules related to generative AI." "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China.

[75] "developing a companion resource to the AI Risk Management Framework, NIST AI 100-1, for generative AI. . . " "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[76] "The China Administration of Artificial Intelligence (CAAI) is the competent authority in charge of the development and administration of artificial intelligence nationwide under the leadership of the central artificial intelligence institution. Other relevant departments and relevant military departments shall, in accordance with the provisions of this Law and applicable laws and administrative regulations, closely cooperate, strengthen coordination, and adequately carry out related work in accordance with the law." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences, April 2024, https://doi.org/10.5281/zenodo.10974163.

[77] ". . . establish at least four new National AI Research Institutes. . . establish an office to coordinate development of AI and other critical and emerging technologies across Department of Energy programs and the 17 National Laboratories. . . establish an HHS AI Task Force that shall. . . develop a strategic plan that includes policies and frameworks — possibly including regulatory

- In China, only one of the two AI law proposals written by legal experts from the Chinese Academy of Social Sciences and industry experts has called for the creation of the China Administration of AI, which would be an overall central authority, sitting under a leading small group on AI (a party institution).[78] The other AI law proposal in China, does not call for the creation of a new institution but rather working through existing institutions. More recently, however, the Third Plenum decision in China has called for instituting new oversight mechanisms related to safety, but it is unclear if this will be a new institution or part of existing institutions.

Technical solutions: Developing technical solutions as a strategy to govern particular risks (e.g., content labelling and watermarking)

- Documents on both the US and Chinese side focus on the use of technical solutions as a strategy to govern particular AI risks.
- There exists significant overlap between the two countries around the need for content provenance and labeling to ensure information traceability. The Biden AI EO emphasizes capturing AI models and their dependencies in "software bills of materials",[79] and US corporate documents have pledged to develop labeling mechanisms for AI-generated content.[80] Similarly, Chinese AI law proposals, amongst other standards and laws, mention the need for AI providers to add implicit identifiers and develop traceability mechanisms for users.[81]
- On the US side, there is a greater focus on privacy-enhancing technologies (PETs) than on the Chinese side. The White House EO refers to PETs being critical to protecting users' privacy and combating broader legal and societal risks that result

action, as appropriate — on responsible deployment and use of AI and AI-enabled technologies in the health and human services sector (including research and discovery, drug and device safety, healthcare delivery and financing, and public health), and identify appropriate guidance and resources to promote that deployment. . . fund the creation of a Research Coordination Network (RCN) dedicated to advancing privacy research and, in particular, the development, deployment, and scaling of PETs." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[78] "The National Artificial Intelligence Office, under the leadership of the central artificial intelligence leadership agency, is responsible for the development and management of artificial intelligence nationwide. Other relevant departments and relevant military departments shall continue to comply with the provisions of this Law, relevant laws and administrative regulations, cooperate closely, strengthen coordination, and complete relevant work in accordance with the law." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences.

[79] "The AI engineering community must institute vulnerability identifiers like Common Vulnerabilities and Exposures (CVE) IDs. Since AI is software, AI models – and their dependencies, including data – should be captured in software bills of materials. The AI system should also respect fundamental privacy principles by default." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[80] "The companies commit to developing robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system. This action enables creativity with AI to flourish but reduces the dangers of fraud and deception." The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

[81] "Artificial intelligence providers should add implicit identifiers in reasonable locations and areas of product and service content, and establish an information traceability mechanism for implicit identifiers to ensure the readability and security of implicit identifiers. If artificial intelligence products and services may cause confusion or misunderstanding by the public, the provider shall take technical measures to add explicit signs that do not affect the use of users in reasonable locations and areas of the product and service content, and remind the public of artificial intelligence in a conspicuous manner." "Scholars' Draft Law on AI," *China Law Society.*

from the improper collection and use of people's data.[82] In Chinese documents we found less extensive mention of PETs, though an Alibaba Whitepaper did argue that technical mitigations should be actively adopted to reduce the collection of personal information.[83]

Evaluations: Requiring, or encouraging, the systematic evaluation of AI systems

- There is some basic agreement for when evaluations are necessary but the content of evaluations are different across the US and China.
- Both China and the US agree on the need for both pre-deployment and post-deployment evaluations and monitoring. The Biden AI EO refers to the need for the government to provide guidance on benchmarks and evaluations, both with respect to pre-deployment testing and post-deployment performance monitoring to ensure that systems function as intended and are resilient to misuse.[84] The NIST AI RMF also refers to the need for implementing post-deployment monitoring plans that capture and evaluate "input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management."[85] Moreover, the FTC's Advice on AI Claims suggests that the FTC's investigations can also act as a form of post-deployment monitoring and intervention.[86] In China, the TC260 Basic Safety Requirements for Generative Artificial Intelligence Services refer to the need for monitoring and evaluation that discover safety issues in the process of service provision and also generally stipulate a range of pre-deployment evaluations that service providers must conduct.[87]
- Both sides also agree on the need for standardization of the evaluations that are run. The Biden AI EO emphasizes the need for 'robust, reliable, repeatable and standard-

[82] "Agencies shall use available policy and technical tools, including privacy-enhancing technologies (PETs) where appropriate, to protect privacy and to combat the broader legal and societal risks — including the chilling of First Amendment rights — that result from the improper collection and use of people's data." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[83] "In the case where generative AI does not have high requirements for personalization, technological means should be actively adopted to lower the collection of personal information from the source and reduce the proportion and authenticity of personal information in the training stage." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence." "

[84] "Testing and evaluations, including post-deployment performance monitoring, will help ensure that AI systems function as intended, are resilient against misuse or dangerous modifications, are ethically developed and operated in a secure manner, and are compliant with applicable Federal laws and policies." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[85] "Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management." National Institute of Standards and Technology U.S. Department of Commerce, January 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[86] "In an investigation, FTC technologists and others can look under the hood and analyze other materials to see if what's inside matches up with your claims." Michael Atleson "Keep your AI claims in check," Federal Trade Commission, February 2023, https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

[87] "Regularized measures for monitoring and evaluation shall be established. Safety issues in the process of service provision discovered by monitoring and evaluation shall be dealt with in a timely manner and the model optimized through targeted instruction fine-tuning and reinforcement learning." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

ized evaluations of AI'.[88] The Standardization Administration of China calls for standards to be developed for 'technical requirements and evaluation methods for artificial intelligence'.[89]

- However, the content of evaluations appears different, though with increasing convergence. The Biden AI EO makes clear the need for evaluations of concerning capabilities such as CBRN threats and cyber capabilities.[90] As mentioned in the dangerous capabilities discussion in the risks section above, there is increasing awareness of such risks from AI in China (e.g., via mention of this in the TC260 GenAI Safety Requirements), but the closest evaluations run with concerning or dangerous capabilities in mind are a safety benchmark by the CAICT, which includes hazardous chemicals as a risk under 'content security' (renamed to bottom lines/red lines in a later update).[91] That said, Chinese companies are aware of the model evaluations being conducted in the West, as these are referenced in the Tencent AI Governance Whitepaper.[92] Baichuan includes alignment and capabilities evaluations, though not dangerous capabilities, in their technical papers.[93][94]

- Interestingly, the US also makes greater mention of the use of AI for evaluations and the defensive advantage it confers than China does. The CBRN threats assessment that the Biden AI EO calls on the Department of Homeland Security to conduct explicitly calls for consideration of the defensive advantages AI confers.[95]

- Finally, the US also appears interested in the continued monitoring of algorithmic

[88] "Meeting this goal requires robust, reliable, repeatable, and standardized evaluations of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[89] "Combined with the actual needs of artificial intelligence governance, standardize the technical research and development and operation services of artificial intelligence, including technical requirements and evaluation methods for artificial intelligence robustness, reliability, traceability, artificial intelligence governance support technology..." "Guidelines for the Construction of a National Comprehensive Standardization System for the Artificial Intelligence Industry," Ministry of Industry and Information Technology.

[90] "Determine the set of technical conditions for a large AI model to have potential capabilities that could be used in malicious cyber-enabled activity, and revise that determination as necessary and appropriate." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210); "...Evaluate the potential for AI to be misused to enable the development or production of CBRN threats..." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[91] "After randomly selecting test questions from a total set of 400,000 Chinese-language questions, CAICT's AI Safety Benchmark conducts a mix of automated and manual reviews of model responses in three major categories: technology ethics, data security, and content security." Jeffrey Ding, "ChinAI #261: First results from CAICT's AI Safety Benchmark."

[92] "Therefore, model evaluation becomes increasingly important. Leading AI companies such as OpenAI and Anthropic are evaluating their AI systems by themselves or in cooperation with external third-party organizations." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[93] "In this section, we report the zero-shot or few-shot results of the pre-trained base models on standard benchmarks. We evaluate Baichuan 2 on free-form generation tasks and multiple-choice tasks." Baichuan Inc., "Baichuan 2: Open Large-scale Language Models."

[94] Another key difference brought to our attention by reviewers, but not present in our dataset, is with respect to evaluation methodology. Evaluations in China tend to be much more static, in the form of simpler benchmarks. Evaluations on dangerous capabilities in particular in the US tend to involve much more dynamic methodologies. For example, agent evaluations and human uplift studies. For a comprehensive characterization of evaluations in China, see: Concordia, "China's AI Safety Evaluations Ecosystem," *AI Safety in China,* September 2024, https://aisafetychina.substack.com/p/chinas-ai-safety-evaluations-ecosystem

[95] "...evaluate the potential for AI to be misused to enable the development or production of CBRN threats, while also considering the benefits and application of AI to counter these threats..." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

decision-making, monitoring it especially for bias. This is brought up by the Equal Employment Opportunity Committee and the Biden AI EO.[96]

Adversarial testing: Evaluation in which the evaluator takes an adversarial approach, seeking to subvert or otherwise produce undesirable results from the system or process being tested by any means available. Sometimes called "red teaming."

- There is some similarity in the adversarial testing required and conducted in both the US and China. In the US, however, the focus is in large part on dual-use foundation models, whilst in China, the requirements apply generally to all kinds of AI systems.
- In both the US and China there is a government requirement for adversarial testing and red-teaming to take place. The Biden AI EO calls for relevant agencies to establish guidelines to enable developers of AI, especially dual-use foundation models, to conduct AI-red teaming tests.[97] In China, the onus appears to be on AI service providers, as a machine learning security assessment standard (GB/T 42888-2023) calls on service providers to conduct adversarial testing across many scenarios ranging from situations where attackers have only access to model inputs and outputs (black-box) to where they have full control of the model internals (white-box).[98]
- There is voluntary industry self-governance with respect to adversarial testing and red-teaming in both jurisdictions as well. The GPT-4, Claude 3, and Gemini technical reports refer to red-teaming and adversarial testing for vulnerabilities, social harms and dangerous capabilities.[99] In China, a Tencent whitepaper covering safety and security

[96] "Employers that are deciding whether to rely on a software vendor to develop or administer an algorithmic decision-making tool may want to ask the vendor, at a minimum, whether steps have been taken to evaluate whether use of the tool causes a substantially lower selection rate for individuals with a characteristic protected by Title VII." "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," U.S. Equal Employment Opportunity Commission; "To address discrimination and biases against protected groups in housing markets and consumer financial markets... require their respective regulated entities, where possible, to use appropriate methodologies including AI tools to ensure compliance with Federal law and evaluate their underwriting models for bias or disparities affecting protected groups;" "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[97] "Establish appropriate guidelines (except for AI used as a component of a national security system), including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[98] "Digital world anti-attack tests should be conducted to test the algorithm's resistance to black-box, white-box, and gray-box attacks; Physical world anti-attack tests should be carried out if conditions permit... Black-box attacks refer to attacks launched by attackers who can only obtain the input and output of the algorithm but do not possess the code, model, or other information. White-box attacks refer to attacks launched by attackers who have full control over the algorithm's inputs and outputs, code, model, and other information." "Information Security Technology – Assessment Specification for Security of Machine Learning Algorithms," Standardization Administration of China.

[99] "Adversarial Testing via Domain Experts... To understand the extent of these risks, we engaged over 50 experts from domains such as long-term AI alignment risks, cybersecurity, biorisk, and international security to adversarially test the model." OpenAI, "GPT-4 Technical Report"; " We apply state-of-the-art red teaming, a form of adversarial testing where adversaries launch an attack on an AI system, in order to test post-trained Gemini models for a range of vulnerabilities (e.g., cybersecurity) and social harms as defined in the safety policies." Gemini Team Google, "Gemini: A Family of Highly Capable Multimodal Models"; "Our RSP requires that we conduct regular risk assessments of our models – primarily through automated evaluations and red teaming – and assign an overall risk level (ASL). We currently evaluate models for three potential sources of catastrophic risk: biological capabilities, cyber capabilities, and autonomous replication and adaption (ARA) capabilities." Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku."

mentions unique risks of AI systems, which they manage in part through an automated prompt security platform that simulates a range of adversarial attacks.[100] They also acknowledge the need for red teams to include a diversity of attack methods and risk scenarios in their evaluations of models.[101]

External auditing: Evaluation by a disinterested counterparty or third party, such as a customer or a professional auditing firm.

- There appears to be some overlap here although 3rd party involvement is seen as favorable or desired in the US, but as just another option service providers can use in China, if they lack internal capacity. Note that as with the rest of the report, this represents facts that could be distilled from written sources and thus may not fully capture undocumented developments or features of the system.
- In the US, through the White House voluntary commitments, companies commit to external security testing, independent evaluations and facilitating third-party discovery and disclosure of security vulnerabilities.[102]
- In China, on the other hand, an Alibaba whitepaper, which states that a "multidimensional security evaluation" should be carried out before use, and that if an AI service provider does not have the capacity to do this themselves, they can use "a neutral third party organization."[103] In addition, a draft law written by a group of scholars from Northwest University for Politics and Law says that "developers and providers of critical artificial intelligence should conduct AI security risk assessments on critical artificial intelligence at least once a year on their own or entrust third-party agencies

[100] "The rapid rise of large model applications has introduced... prompt risks, including prompt injection and adversarial attacks. In response to the new security risks unique to such large models, we have built a prompt security evaluation platform, which is specially used to simulate the behavior of attackers to understand the security and performance of large models in risk scenarios associated with prompts... assists the business in reducing risk during the process of launching the large model to ensure that its response content complies with various laws and regulations such as the 'Interim Measures for the Management of Generative AI Services'. Therefore, prompt security assessment requires automated attack sample generation and automated risk analysis capabilities." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[101] "The second is adversarial testing or red teaming. In short, before the model is released, internal or external professionals are invited to serve as white hat hackers and launch various adversarial attacks on the model in the red-team tests to evaluate the product's security measures and ability to withstand external attacks, so as to identify potential problems and resolve them." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

[102] "The companies commit to internal and external security testing of their AI systems before their release. This testing, which will be carried out in part by independent experts, guards against some of the most significant sources of AI risks, such as biosecurity and cybersecurity, as well as its broader societal effects." The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI"; "The companies commit to facilitating third-party discovery and reporting of vulnerabilities in their AI systems. Some issues may persist even after an AI system is released and a robust reporting mechanism enables them to be found and fixed quickly." The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

[103] "Model verification: The service provider verifies the model and completes multidimensional security evaluation before use. The service provider may not necessarily have the ability to carry out multidimensional security evaluation, and evaluation services need to be provided by a neutral third party organization." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence.".

to conduct timely rectification of discovered security issues and report to the artificial intelligence authorities".[104]

Licensing or registration: Requiring, incentivizing, or otherwise encouraging actors involved in AI-related activities, such as AI developers, vendors, users, or researchers, to either receive sanction from a regulator for their activities (licensing) or to notify a regulator of their activity pursuant to a formal process (registration).

- There is moderate overlap in this category as both jurisdictions are increasingly focused on tiered oversight for more capable AI models implemented through some type of licensing or registration system.
- The US has notification and reporting requirements (similar in nature to what we define as registration) extending to models that are trained above a certain threshold (10^{26} FLOPs for general models, 10^{23} FLOPs for models trained on biological sequence data).[105] The Biden AI EO also calls for new reporting requirements on cloud compute usage by foreign actors training AI through American Infrastructure as a Service (IaaS) Providers.[106] A separate executive order calls for creation of regulation wherein US persons will need to provide notification of information relative to specific transactions that might be related to investments in national security technologies.[107]
- Currently in China there is a licensing system, wherein service providers using algorithms with social mobilization properties used in recommender systems, deep synthesis and generative AI systems, must file their algorithms, alongside security assessments and other documents with the Cyberspace Administration of China.[108]

[104] "Developers and providers of critical artificial intelligence should conduct artificial intelligence security risk assessments on critical artificial intelligence at least once a year on their own or entrust third-party agencies to conduct timely rectification of discovered security issues and report to the artificial intelligence authorities." "Scholars' Draft Law on AI," *China Law Society.*

[105] Until such technical conditions are defined, the Secretary shall require compliance with these reporting requirements for: (i) any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 10^{23} integer or floating-point operations; and (ii) any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 10^{20} integer or floating-point operations per second for training AI." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[106] "Propose regulations that require United States IaaS Providers to submit a report to the Secretary of Commerce when a foreign person transacts with that United States IaaS Provider to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity (a "training run"). Such reports shall include, at a minimum, the identity of the foreign person and the existence of any training run of an AI model meeting the criteria set forth in this section..." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[107] " Section 1. Notifiable and Prohibited Transactions. (a) ... issue, subject to public notice and comment, regulations that require United States persons to provide notification of information relative to certain transactions involving covered foreign persons (notifiable transactions) and that prohibit United States persons from engaging in certain other transactions involving covered foreign persons (prohibited transactions). (b) The regulations issued under this section shall identify categories of notifiable transactions that involve covered national security technologies and products that the Secretary... determines may contribute to the threat to the national security of the United States identified in this order. The regulations shall require United States persons to notify the Department of the Treasury of each such transaction..." "Executive Order 14105: Addressing United States Investments in Certain National Security Technologies and Products in Countries of Concern," Federal Register 88 (154).

[108] "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," Cyberspace Administration of China.

- AI law proposals from legal scholars and industry in China call for potential registration or licensing oversight regimes, depending on the capability or compute threshold of models, as well as their potential national security or economic impact.[109]

3.3.3 Weak overlap

Risk assessments: Identification, assessment and in some cases quantification of potential sources of and pathways to harm caused by AI systems.

- There are various risk assessment frameworks used across the US and China. There seems to be limited clear overlap, although most mentions are broad enough to potentially encompass many similar ideas.
- In the US, the Biden AI EO encourages sector-specific risk assessments, and also assessment of how current policy tools can assist in AI-related disruptions to the labor market.[110] The NIST RMF states that risk management can be enhanced by tracking emergent risks and considering techniques for measuring them.[111]
- In China, the dominant language around risk assessments comes from ethics reviews, which accords with the broader push for science and technology ethics management practices in the country. MIIT calls for the development of standards around ethical risk assessments for AI,[112] while Sensetime, in a whitepaper, lay out how they grade ethics risks levels in 5 tiers, based on the impact of a product to 'product safety, personal rights, market fairness, public safety and environmental health'.[113]

Disclosure: Requiring or encouraging, the disclosure of information about AI systems by their users, developers, vendors, or other stakeholders to third parties, including but not

[109] "The State establishes an Artificial Intelligence Negative List, subjecting products and services within the negative list to a licensing oversight system and, where necessary, those outside the negative list to a registry oversight system." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences; "Key artificial intelligence providers should register through the national artificial intelligence supervision platform and complete the filing procedures within 7 working days from the date of receipt of the certification notice." "Scholars' Draft Law on AI," China Law Society.

[110] "Independent regulatory agencies are encouraged, as they deem appropriate, to contribute to sector-specific risk assessments." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[111] "Organizations' risk management efforts will be enhanced by identifying and tracking emergent risks and considering techniques for measuring them. AI system impact assessment approaches can help AI actors understand potential impacts or harms within specific contexts." National Institute of Standards and Technology U.S. Department of Commerce, January 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[112] "...standardize the technical research and development and operational service requirements of artificial intelligence, including technical requirements and evaluation methods for artificial intelligence robustness, reliability, traceability, artificial intelligence governance support technology; standardize the entire life of artificial intelligence." "Guidelines for the Construction of a National Comprehensive Standardization System for the Artificial Intelligence Industry," Ministry of Industry and Information Technology.

[113] "At the application level, we have established an Ethics Risk Classification Management Mechanism and an Ethics Risk Review Team to carry out graded and targeted ethics risk management throughout the entire product life cycle of design, development, deployment, and operation. We have also set up supporting processes for self-inspection, assessment and review of risks, and follow-up reviews. We classify ethics risks from low to high... based on the impact of final product safety, personal rights and interests, market fairness, public safety, and environmental health." "AI Governance Whitepaper," SenseTime.

limited to the general public.

- While both countries agree on the need to disclose information about AI systems, they focus on different areas and implement varying degrees of enforceability.
- The US prioritizes national security-related issues, through a focus on dual-use foundation models and malicious cyber activities (as specified in the Biden AI EO).[114] As a result, the federal government is the primary entity that companies must report to. For instance, companies developing potential dual-use foundation models are required to provide the federal government with ongoing information about their models.[115] More broadly, leading companies in the US, through the White House Voluntary commitments, have pledged to share information with a wide range of actors about the mitigation of AI risks, as well as to publicly reporting their system capabilities, limitations, and areas of appropriate or inappropriate use.[116]
- Conversely, China's disclosure requirements focus on transparency towards users, as well as the government. While the algorithm registry is oriented toward disclosure of algorithms and models to the government, companies are also obligated to disclose information such as training data (including notifying individuals whose sensitive data are being used and edited), the purpose and functionality of systems, and other relevant details to their users.[117] Chinese companies have also observed and may potentially model after their American counterparts in terms of reporting vulnerabilities. In its Large Model Security and Safety Report, for example, Tencent pointed to American companies' voluntary agreements on AI.[118]

[114] "Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records regarding the following: any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats; the ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights;" "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[115] Ibid.

[116] "The companies commit to sharing information across the industry and with governments, civil society, and academia on managing AI risks. This includes best practices for safety, information on attempts to circumvent safeguards, and technical collaboration." The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI."

[117] "Where deep synthesis service providers and technical support provide functions for editing biometric information such as faces and voices, they shall prompt the users of the deep synthesis service to notify the individuals whose personal information is being edited and obtain their independent consent in accordance with law." "Provisions on the Administration of Deep Synthesis Internet Information Services," Cyberspace Administration of China; "Service transparency: 1) If the service is provided using an interactive interface, information such as the people, situations, and uses for which the service is suitable shall be disclosed to the public in a prominent location. . . 2) If the service is provided using an interactive interface, the following information shall be disclosed to the users on the homepage of the website, the service agreement, and other easily viewed locations: — Limitations of the service; — Summary information on the models and algorithms used, etc.; — The personal information collected and its uses in the service." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[118] "Sixth, share information on AI risks. The leading AI companies in the United States made a voluntary commitment to the White House administration: to promptly share information about AI risks with the government, society, and academia, such as sharing risk information and test results with the U.S. government before launching AI products, promptly disclosing model limitations to the public, etc." Tencent Research Institute (TRI), Tencent Zhuque lab, Tencent Hunyuan model team, Tsinghua Shenzhen International Graduate School, and Zhejiang University State Key Lab of Blockchain and Data Security, "Tencent Large Model Security and Safety Report."

- Another difference between the two countries' disclosure requirements is their degree of codification. Disclosure in China has a longer history stretching back to at least early 2022 with the algorithm registry process,[119] whereas disclosure via the Defence Production Act (DPA) in the US is much newer and less durably established via an executive order rather than legislation.[120]

Input controls: Restricting or placing conditions on the sale, distribution, or use of technical inputs to AI systems, specifically data or computational resources.

- Both the US and China refer to compute and data related controls and restrictions, though with significant differences.
- In terms of compute, the Biden AI EO makes clear that notification and reporting requirements extend to models that are trained above a certain threshold (10^{26} FLOPs for general models, 10^{23} FLOPs for biological sequence data).[121] In China, the two AI law proposals from academic experts both refer to the use of compute thresholds in terms of defining models or AI systems subject to greater control, but without specifying an exact threshold.[122] The Biden AI EO extends notification requirements to organizations that possess a computing cluster with a theoretical maximum capacity of 10^{20} FLOP/s,[123] something that is not mentioned in any Chinese documents. The Biden AI EO also calls for new reporting requirements on cloud compute usage by foreign actors training AI through American Infrastructure as a Service Providers.[124]

[119] "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," Cyberspace Administration of China.

[120] "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[121] Until such technical conditions are defined, the Secretary shall require compliance with these reporting requirements for: (i) any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 10^{23} integer or floating-point operations; and (ii) any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 10^{20} integer or floating-point operations per second for training AI." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[122] "Key artificial intelligence includes the following types. . . a basic model that reaches a certain level in terms of computing power, parameters, usage scale." "Scholars' Draft Law on AI," China Law Society; "Foundation Models refer to artificial intelligence models that have undergone training with the accumulation of computing power investment to a certain scale, serving general purposes, and capable of providing technological support for a wide range of downstream services. The floating-point operations (FLOPs) and other computing power standards for the identification of foundation models shall be formulated, publicly issued, and regularly updated by the National AI Administrative Authority." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences.

[123] Until such technical conditions are defined, the Secretary shall require compliance with these reporting requirements for: (i) any model that was trained using a quantity of computing power greater than 1026 integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 1023 integer or floating-point operations; and (ii) any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 1020 integer or floating-point operations per second for training AI." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[124] "Propose regulations that require United States IaaS Providers to submit a report to the Secretary of Commerce when a foreign person transacts with that United States IaaS Provider to train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity (a "training run"). Such reports shall include, at a minimum, the identity of the foreign person and the existence of any training run of an AI model meeting the criteria set forth in this section. . . " "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

- In terms of data, this is mostly a focus of Chinese sources in our dataset. The TC260 Basic Safety Requirements for Generative Artificial Intelligence Services refer to the need to filter training datasets to remove "illegal and unhealthy information",[125] whilst an Alibaba whitepaper focuses on the need to manage the IP infringement issues of training datasets by purchasing databases from real rights holders.[126] The two AI law proposals mentioned above also refer to data controls, specifically referencing training data and output data filtering to reduce bias and discrimination.[127]
- *Given our focus on domestic-facing regulations, we have excluded the US semiconductor export controls from our study.*

Liability: Holding specific parties accountable through civil or criminal penalties for unlawful actions.

- Generally, both countries state that AI developers need to comply with existing laws, but what they focus on is quite different.
- In the US, there are two main thrusts to liability. A softer form of pseudo-liability is mentioned but not enacted through language referring to the accountability of those deploying AI systems in the justice system and federal government to protect affected communities against bias and discrimination, in the preamble to the AI EO.[128] A second harder and clearer area of liability is in reference to consumer protection law, and mentioned both in the Biden AI EO and the FTC's guidance on use of AI in deception.[129] Here the focus is on preventing fraud, with the FTC stating that it will go after companies that use AI in deceptive practices.[130]

[125] "Filtering of training corpus content: Methods such as keywords, classification models, and manual sampling inspection shall be adopted to thoroughly filter out all illegal and unhealthy information in corpora." "Basic Safety Requirements for Generative Artificial Intelligence Services," National Technical Committee 260 on Cybersecurity of Standardization Administration of China.

[126] "The key point to reduce intellectual property infringement caused by Generative AI is before the formation of the training data set. The common solutions include: (1) to purchase databases with intellectual property rights from real rights holders (2) to use legally authorized open source data sets (3) to avoid crawling beyond technical measures." Alibaba AI Governance Research Center, "White Paper on the Governance and Use of Generative Artificial Intelligence."

[127] "Article 36 (Fairness Obligation) Artificial Intelligence developers should take necessary measures to effectively prevent harmful bias and discrimination during the process of training data processing and annotation, algorithm model design, development, and verification testing. Artificial intelligence providers should strengthen the management of input data and output data in the process of providing products and services to effectively prevent harmful prejudice and discrimination." "Artificial Intelligence Law Model Law Version 1.1 (Expert Suggestion Draft)," Chinese Academy of Social Sciences, September 2023, https://www.21jingji.com/article/20230907/herald/982ae3bb7b82597b4dc1f990ded64ad2.html; "AI developers and providers shall carry out a safety risk assessment before providing products and services, and record the handling circumstances. The AI safety risk assessment shall include the following: (i) Whether there is potential bias or discrimination;" Zhang Linghan et al., Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars), March 2024, https://perma.cc/L9E4-5K3V (Translated by CSET, May 2024, https://cset.georgetown.edu/publication/china-ai-law-draft/).

[128] "My Administration will build on the important steps that have already been taken... in seeking to ensure that AI complies with all Federal laws and to promote robust technical evaluations, careful oversight, engagement with affected communities, and rigorous regulation. It is necessary to hold those developing and deploying AI accountable to standards that protect against unlawful discrimination and abuse, including in the justice system and the Federal Government." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[129] "The Federal Government will enforce existing consumer protection laws and principles and enact appropriate safeguards against fraud, unintended bias, discrimination, infringements on privacy, and other harms from AI." "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[130] "The FTC Act's prohibition on deceptive or unfair conduct can apply if you make, sell, or use a tool that is effectively

- In China, regulations on AI, including the recommendation algorithm and interim generative AI measures, hold service providers liable in at least two ways. Where personal information is involved, the providers bear responsibility as information handlers.[131] Moreover, if algorithms are not filed with the relevant authorities or if corrections requested by authorities are refused, providers can be fined or have their services suspended.[132]

- Chinese AI law proposals by experts suggest further extensions of liability, through liability for misuse and failure to carry out adequate risk prevention in the CUPL draft law,[133] and failure to adhere to assigned responsibilities in the CASS draft law.[134] The CUPL draft law also shifts some liability towards downstream users, in stating that foundation model developers are not responsible as long as they agree on terms with downstream users on how to use the models safely. The responsibility thus rests with those who use a foundation model, either at the application, fine-tuning or user level.[135] We see some evidence of this, as platforms such as Douyin (a Chinese domestic version of Tiktok), specify that publishers on Douyin are responsible for the consequences of the content that they generate.[136]

designed to deceive – even if that's not its intended or sole purpose." Michael Atleson, "Chatbots, deepfakes, and voice clones: AI deception for sale," Federal Trade Commission, March 2024, https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale.

[131] "Providers shall bear responsibility as the producers of online information content in accordance with law and are to fulfill the online information security obligations. Where personal information is involved, they are to bear responsibility as personal information handlers and fulfill obligations to protect personal information." "Interim Measures for the Management of Generative Artificial Intelligence Services," Cyberspace Administration of China.

[132] "If an AI provider is required to file a registry but fails to do so, the National AI Administrative Authority shall issue a warning. In severe cases, a fine between ten thousand and one hundred thousand yuan may be imposed. If an AI provider obtains registry through improper means, such as concealing relevant information or providing false materials, the National AI Administrative Authority shall revoke the registry, issue a warning, and make a public criticism. In severe cases, a fine between one hundred thousand and one million yuan may be imposed. If an AI provider terminates its service without going through the procedures to cancel the registry, or if it is subjected to administrative penalties such as being ordered to shut down the website, having its relevant business permit revoked, or having its operation license revoked due to serious legal violations, the National AI Administrative Authority shall cancel the registry." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences.

[133] "Where critical AI products and services cause damages to others and the provider cannot prove that it is not at fault, the provider shall bear tort liability." "Scholars' Draft Law on AI," China Law Society, March 2024, http://www.fxcxw.org.cn/dyna/content.php?id=26910.

[134] "If an AI provider is required to file a registry but fails to do so, the National AI Administrative Authority shall issue a warning. In severe cases, a fine between ten thousand and one hundred thousand yuan may be imposed. If an AI provider obtains registry through improper means, such as concealing relevant information or providing false materials, the National AI Administrative Authority shall revoke the registry, issue a warning, and make a public criticism. In severe cases, a fine between one hundred thousand and one million yuan may be imposed. If an AI provider terminates its service without going through the procedures to cancel the registry, or if it is subjected to administrative penalties such as being ordered to shut down the website, having its relevant business permit revoked, or having its operation license revoked due to serious legal violations, the National AI Administrative Authority shall cancel the registry." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences.

[135] "If the use of AI products and services causes damages to others and the user is at fault, the user shall bear tort liability; if the developer or provider of the AI has failed to fulfill its obligations under this Law, it shall bear the corresponding tort liability. Where [other] laws dictate otherwise, liability shall be attributed in accordance with their provisions." "Scholars' Draft Law on AI," China Law Society, March 2024, http://www.fxcxw.org.cn/dyna/content.php?id=26910.

[136] "5. Publishers are responsible for the consequences of content generated by artificial intelligence, regardless of how the content is generated." "Douyin's Platform Norms and Industry Initiatives on Content Generated by Artificial Intelligence," Douyin, May 2023, http://www.cm3721.com/m/view.php?aid=29171 (Translated by China Law Translate, May 2023, https://www.chinalawtranslate.com/en/dou-yin-ai-rules/)

Pilots and testbeds: Creating, facilitating, setting conditions on, or otherwise addressing the development and operation of government-supported or government-conducted pilot programs or test environments related to artificial intelligence.

- In the US, testbeds refer to a facility or mechanism for the rigorous, transparent and replicable testing of tools. The Biden AI EO calls for the development of testing environments that support the development of privacy-enhancing technologies, implementing a plan to developing the Department of Energy's model evaluation tools and testbeds, as well as working with partners to leverage the Department of Energy's testbeds to create foundation models that support new applications in science and energy.[137]
- Amongst our Chinese documents, we saw much more limited mention, with only a Chinese draft law by academic experts calling for state agencies to pilot the use of AI in public services.[138]

[137] "developing and helping to ensure the availability of testing environments, such as testbeds, to support the development of safe, secure, and trustworthy AI technologies, as well as to support the design, development, and deployment of associated PETs. . . develop and. . . implement a plan for developing the Department of Energy's AI model evaluation tools and AI testbeds. . . (iv) take steps to expand partnerships with industry, academia, other agencies, and international allies and partners to utilize the Department of Energy's computing capabilities and AI testbeds to build foundation models that support new applications in science and energy, and for national security, including partnerships that increase community preparedness for climate-related risks, enable clean-energy deployment (including addressing delays in permitting reviews), and enhance grid reliability and resilience;" "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register 88 (210).

[138] "Governmental agencies, public institutions, and other organizations legally endowed with the function of managing public affairs are encouraged to pioneer and pilot the application of artificial intelligence technology in fields such as government services and public management in accordance with the law, giving priority to the procurement and use of safe and reliable artificial intelligence products and services." "The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version," Chinese Academy of Social Sciences.

4 Recommendations

Our results show that American and Chinese actors share some concerns and concepts in AI safety and governance—as expressed in their own words, declarations, and policy proposals.

In this section, we discuss our key results, **focusing on issues where we found moderate and strong overlap**. For each of the issues we analyze in this section, we compare the areas of overlap surfaced in our analysis against the broader context of the US–China relationship, as well as existing international dialogue on AI, to determine whether dialogue on a given topic could be effective.

First, we discuss the strong and consequential common ground we observe on issues related to AI risk.

Next, we discuss the degree to which we observed overlap on governance approaches. Acknowledging common ground here, while working to overcome divergence, could prove especially critical for international efforts to address global AI challenges, prevent regulatory arbitrage, and reap the benefits of interoperability.[139]

Finally, we synthesize the common ground found across issues related to AI risk and governance approaches into a set of recommendations. Specifically, we combine related issues across our analysis of risks and governance approaches and recommend areas where new dialogues or further investment in existing efforts may be productive.

4.1 Risks

The field of AI governance and safety is nascent. With much of the world 'waking up' to AI governance after the release of ChatGPT in 2023, we believed it was necessary to not focus too only on what the US and China were doing with respect to AI governance, but also how they were talking about AI risks.

With respect to AI risks, we find that the US and China have significant common ground—virtually all of the risks considered in our dataset have either strong or moderate overlap—suggesting that for almost all of the risks, there is at least some shared understanding. For two risks in particular—limited user transparency and poor reliability—there are similar concerns demonstrated by the sources we analyzed.

Yet, there is virtually no dialogue between the US and China on almost any of these shared understandings of risks. Part of this can be explained by the risk of issue linkage to broader sources of tension in the US–China relationship, such as cybersecurity. However, the critical

[139] Duncan Cass-Beggs et al., "Framework Convention on Global AI Challenges," *Centre for International Governance Innovation*, June 2024, https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/; Claire Dennis et. al, "What Should be Internationalized in AI Governance," *Oxford Martin School AI Governance Initiative,* November 2024, https://oms-www.files.svdcdn.com/production/downloads/What%20should%20be%20internationalised%20in%20AI%20Governance-final.pdf?dm=1731486256

takeaway from our analysis is that there is space for *more* dialogue between the US and China, both by strengthening existing dialogue and by starting new forms of dialogue.

Specifically, US and Chinese actors can consider:

1. **Strengthening existing state-level cooperation on managing the risks from dangerous capabilities of AI systems, and limiting proliferation of dangerous AI systems by protecting model weights.** Both countries have shown an interest in managing the risks from dangerous national security-relevant capabilities that advanced AI models may possess and in preventing theft of model weights. Given the overlap on these topics found in our analysis, we believe that more dialogue on these topics is possible, especially on preventing theft of model weights by non-state actors.
2. **Starting technical standards-setting discussions on reliability and robustness.** Both Chinese and US sources demonstrated similar understanding of the risks and interest in addressing these issues. At the same time, these issues have not been a significant part of ongoing international discussions, most of which have taken place at a very high-level. These issues are likely best addressed at the level of technical standards. For example, both countries could cooperate through existing standards-setting organizations such as the ISO or IEC on these topics, or set up working-level government-to-government technical cooperation.

The table across the next few pages contains a more detailed discussion of the risks that had strong or moderate overlap in our analysis, alongside our assessment of whether such overlap could lead to productive dialogue.

🟩 Strong overlap
Moderate overlap

Table 4: Analysis of AI Risk Overlap Between US and China

Risks	Existing US–China context	Similarities surfaced	Space for further dialogue
🟩 Limited user transparency	Not a significant part of ongoing international dialogues or discussions, briefly mentioned in the Bletchley declaration.	US and Chinese documents both understand user transparency in a similar way, defining it as the extent to which information about an AI system and its outputs is available to a user. There is common ground in the importance of ensuring user transparency.	**?** While there is common ground on this issue, it is unclear whether user transparency needs to be discussed at the bilateral or multilateral level.
🟩 Poor reliability	Not a significant part of ongoing international dialogues or discussions	Both NIST in the US and CAICT in China mention reliability as a desired goal for the overall correctness of an AI system.	✓ Dialogue on reliability standards could be a topic for standards-setting bodies to take on given the technical nature of the problem. These discussions could focus on defining reliability for general purpose AI and for the use of AI in specific industries.

Continued on next page

Table 4 – continued from previous page

Risks	**Existing US–China context**	**Similarities surfaced**	**Space for further dialogue**
Lack of robustness	Not a significant part of ongoing international dialogues or discussions	Both American and Chinese sources talk about robustness, generally linking it to the idea of reliability under more adverse or unusual circumstances. However, it is unclear whether the definitions used are exactly the same in each case.	**?** Discussion of robustness could be part of reliability standards-setting discussions mentioned above.
Bias and discrimination	Bias mitigation and reduction mentioned in both the Bletchley declaration and the US-sponsored UNGA resolution. Bias and discrimination are adjacent to existing tensions around politics and values, as the US and China both view each other as discriminating against segments of their domestic populations. There are also likely significant differences in how each side defines discrimination and bias.	US and Chinese documents both have similar stated scope when it comes to bias, and are concerned with integration of unintended bias into decision-making, the need for testing for bias and reducing bias across the entire model lifecycle.	× Given the intersection of bias and discrimination with existing tensions, it’s unlikely that further discussion of bias in systems would be productive.

Continued on next page

Table 4 – continued from previous page

Risks	Existing US–China context	Similarities surfaced	Space for further dialogue
Lack of interpretability and explainability	Mentioned briefly in the Bletchley declaration	While both US and Chinese documents discuss the need to improve the interpretability of AI systems, Chinese documents appear to be less sanguine about the ability to develop interpretable systems compared to US documents, such as the NIST Risk Management Framework.	× Given limited mention of this in existing international discussions and less clear common ground evident from documents in our database, it appears less likely that this would be an easy area for dialogue.
Dangerous capabilities	Mentioned briefly in the Bletchley declaration, including references to CBRN and loss of control issues (i.e. model autonomy). Concerns around the dangerous capabilities of AI systems do appear to be part of existing US–China Track 1 dialogues. Ongoing Track II processes, such as IDAIS do suggest that there is academic consensus, at the very least, on dangerous capabilities.	There is increasing convergence evident between Chinese and American sources; particularly that models can display dangerous capabilities and that these capabilities should be tested for. This overlap is strong in areas such as bio-chemical capabilities and autonomy-related capabilities (e.g., self-replication), and weaker with respect to persuasion as a dangerous capability.	✓ Given the presence of significant common ground at the Track II level and in stated documents, this a promising area for future dialogues. In particular, intergovernmental dialogues could try to build common ground on the domains where dangerous capabilities are most concerning.

Continued on next page

Table 4 – continued from previous page

Risks	Existing US–China context	Similarities surfaced	Space for further dialogue
Weak cybersecurity	Not a significant part of ongoing international dialogues or discussions	Both countries express concerns about weak cybersecurity, which can be roughly categorised into concerns related to theft of model weights, and general concerns about software vulnerabilities. There is much clearer common ground across the Biden AI EO, Chinese corporate lab reports and the TC260 GenAI Safety Standard that theft or undesired access to model weights should be avoided.	✓ While broader US-China relations with respect to cyber vulnerabilities are adversarial, the common ground on protecting model weights from theft and undesired access creates some room for dialogue. Particularly, dialogue could focus on which actors both parties want to protect against (e.g., non-state actors).

4.2 Governance approaches

A necessary complement to understanding how risks from AI are discussed, is considering what is being done to manage risks. We thus focused the second pillar of our analysis on governance approaches.

Unsurprisingly, the overlap found between US and Chinese approaches to governance is significantly weaker than that found on risks. Even complete agreement on what the risks are could lead to diverging approaches on what needs to be done to manage these risks. Moreover, approaches to governance are fast evolving, thus governance approaches that currently diverge may only be diverging temporarily.

The converse may also hold true, as governance approaches that appear to converge today, may only be doing so superficially. International dialogue and cooperation is critical to ensuring that governance approaches remain coordinated, especially when risks can only be managed through international cooperation (e.g., transnational risks like AI-enabled cyber attacks carried out by globally distributed non-state actors).

Specifically, US and Chinese actors can consider cooperation on five governance approaches:

1. **Strengthening cooperation between governments on select aspects of model evaluations, such as agreeing to critical thresholds, which if crossed would signal that AI models pose significant domestic and global national security risks.** While both countries acknowledge a central role for evaluation and testing, there are differences in the content of the evaluations that are most emphasized, which could be a source of potential tension. Nevertheless, given existing consensus on the importance of testing and auditing AI systems, and the common ground established in Track II dialogues such as the International Dialogues on AI Safety, the two governments can agree to a common set of critical thresholds, such as capability or risk-based red lines.

2. **Starting industry-led cooperation on content provenance.** There is common ground between the US and China on the need for technical solutions such as watermarking to improve content provenance. Cooperation on this front can be led by industry associations such as the US-based Coalition for Content Provenance and Authenticity (C2PA) which already focus on this issue.

3. **Starting cooperation on technical standards related to adversarial testing of systems.** Adversarial testing is an important area for both China and the US. Such testing is required to ensure an AI system is robust, and can be incorporated into dialogues covering risks related to reliability and robustness (see the Risks section above).

4. **Sharing best practices on the use of AI for pro-safety purposes at the Track II level.** The use of AI for pro-safety purposes is an area of overlap, as both the US

and China think through how AI could be used to improve the safety of AI systems. Moreover, the non-zero-sum nature of this type of AI could make it a simpler area for dialogue. However, it is worth noting that some sub-topics (e.g., the use of AI to bolster cyberdefense) may be less suitable than others (e.g., the use of AI to score model evaluations).

5. **Strengthening Track II dialogues on new governance tools for AI (e.g., compute thresholds, model registration).** Both American and Chinese policymakers are interested in developing new tools that will allow them to approach the governance of AI in a structured way. For example, compute thresholds are a feature both in existing American regulation (i.e. the executive order on AI) and two Chinese AI law proposals. Track II dialogue focused on best practice sharing on AI governance approaches could be an effective way of ensuring that the latest thinking on AI governance diffuses widely. This would also be critical for future harmonization or interoperability efforts.

The table across the next few pages contains a more detailed discussion of the governance approaches that had strong or moderate overlap in our analysis, alongside our assessment of whether such overlap could lead to useful dialogue.

Strong overlap
Moderate overlap

Table 5: Analysis of AI Governance Approaches Between US and China

Category	**Existing US–China context**	**Similarities surfaced**	**Space for further dialogue**
Use of AI for pro-safety purposes	The US-led UNGA resolution on AI includes references to the potential of AI to accelerate progress on the sustainable development goals.	Both US and Chinese sources acknowledge the role that AI systems can play in pro-safety solutions, such as detecting vulnerabilities in code and in automating AI model evaluations.	✓ While there may be some potential national security sensitivities around sharing how AI is being used to strengthen defenses, this does not preclude the discussion of this broad topic.
Convening	Mentioned in the Bletchley declaration in the context of deepening and broadening international cooperation, as well as the need to involve actors from many different sectors in AI governance.	Both the US and China have explicit language around convening a variety of stakeholders to make progress on AI governance and safety.	× While convening has been mentioned as both an international and domestic governance approach, it is unclear what a dialogue on this topic would focus on.

Continued on next page

Table 5 – continued from previous page

Category	**Existing US–China context**	**Similarities surfaced**	**Space for further dialogue**
Governance development	The Bletchley declaration and US-led UNGA AI Resolution both briefly mention countries developing governance approaches in line with national conditions, while also cooperating internationally. US and Chinese governance systems have clearly been positioned as fundamentally at odds by both leaders, driven by differences in values and political systems, so dialogue around sharing best practices in domestic governance or coordinated international governance, could potentially be complicated by links to the broader divergence in values and systems.	Both the US and China show clear interest in governance development. Both appear interested in collaborating internationally to build shared standards, and developing domestic governance infrastructure. The US has thus far shown a greater interest in setting up new institutions within the executive branch while new institutions are only mentioned in one of two AI law proposals in China.	**?** While there is some basis for a dialogue on this topic linked to existing international documents (the Bletchley declaration and the US-led UNGA AI resolution) and overlap found in our analysis, it is unclear whether carving out governance development as a separate topic would be beneficial or possible. A Track II dialogue focused on sharing innovative governance methods suited for AI (e.g., compute thresholds, model registration) may be a fruitful way to share lessons learnt on AI governance.

Continued on next page

Table 5 – continued from previous page

Category	**Existing US–China context**	**Similarities surfaced**	**Space for further dialogue**
Technical solutions	Technical solutions linked to watermarks and content provenance are mentioned in the US-led UNGA resolution as an example of an international interoperable tool or standard	Documents on both the US and Chinese side focus on the use of technical solutions as a strategy to govern particular AI risks. This is particularly true with regards to content provenance and labeling required to ensure information traceability. There are differences with regards to emphasis on other technical solutions. In the US, the Biden AI EO spotlights privacy-enhancing technologies, which seem to receive less attention in China.	✓ Specific industry-led standards-setting efforts on watermarking and content provenance could prove fruitful given the common ground found in our analysis, and the mention of in the US-led UNGA resolution.

Continued on next page

Table 5 – continued from previous page

Category	**Existing US–China context**	**Similarities surfaced**	**Space for further dialogue**
Evaluations	Tools for evaluation and testing are mentioned in the Bletchley declaration and the US-led UNGA resolution. The need for monitoring of dangerous capabilities in AI models is also mentioned in Track II dialogues such as the International Dialogues on AI Safety, which mention the need to monitor and enforce redlines in AI development.	There is some basic agreement for when evaluations are necessary. Both China and the US also agree on the need for both pre-deployment, post-deployment evaluations and monitoring, as well as the need for standardization of the evaluations that are run. However there are some differences with respect to the content of evaluations. The US places greater emphasis on evaluations of dangerous capabilities (e.g., CBRN risks) while China places greater emphasis on evaluating for politically sensitive content. Nevertheless, there is still some convergence with Chinese actors increasingly interested in evaluations for dangerous capabilities (e.g., a recent safety benchmark by the CAICT includes concerns about hazardous chemicals)	✓ There is a foundation for further dialogue suggested by the Track II processes, the Bletchley declaration and the US-led UNGA AI resolution. There is also clear common ground found in our database, as a wide range of Chinese and American actors appreciate the value of developing a clear empirical understanding of AI models through model evaluation and testing. Furthermore, dialogue on evaluations could naturally be linked to shared concerns about dangerous capabilities. (see the Risks section above)

Continued on next page

Table 5 – continued from previous page

Category	Existing US–China context	Similarities surfaced	Space for further dialogue
Adversarial testing	There is limited mention of this specific approach to evaluations within international dialogue.	There is some similarity in the adversarial testing required and conducted in both the US and China. In the US, however, the focus is in large part on dual-use foundation models, whilst in China, the requirements apply generally to all kinds of AI systems.	✓ Given the focus of adversarial testing on ensuring the robustness and reliability of AI systems, this topic could be combined with a standards-setting process on robustness and reliability. (see the Risks section above)
External auditing	The importance of third party evaluators and vulnerability discovery is mentioned in the Frontier Safety Commitments, which were signed by leading AI companies, including Zhipu, a Chinese frontier AI startup. The need for third-party auditing is also mentioned in the 1st IDAIS consensus statement.	There appears to be some overlap here although third party involvement is seen as favorable or desired in the US, but as just another option service providers can use in China, if they lack internal capacity.	✓ The role of third party auditors and evaluators could be a productive sub-topic within a broader dialogue on evaluations
Licensing or registration	The need for model registration is mentioned in both the 1st and 2nd IDAIS consensus statements.	There is moderate overlap as both jurisdictions are increasingly focused on tiered oversight for more capable AI models implemented through some type of licensing or registration system.	**?** While registration systems are likely to be very different, dialogue on this issue could be part of a broader dialogue of novel governance approaches, likely in a Track II rather than official setting.

4.3 Consolidated recommendations

The consolidated picture that emerges from our analysis is a promising one. With respect to risks, we identify one new opportunity for dialogue and one opportunity for strengthening existing dialogue. Through our analysis of governance approaches we identify three new opportunities for dialogue and two opportunities for strengthening existing dialogue.

In this section, we combine these potential dialogue topics into a more holistic set of recommendations, including suggestions for actors who may be best placed to carry forward these dialogues.

First, we recommend that US and Chinese governments should strengthen existing intergovernmental dialogue, covering issues related to national security, such as evaluating models for 'dangerous capabilities' and preventing proliferation to non-state actors[140].

Both countries emphasize the importance of testing and evaluating AI systems, and are concerned about the dangerous national security-relevant capabilities that AI systems may possess. While there are some differences – Chinese safety/security assessments so far have focused less on dangerous national security-relevant capabilities of AI models and more on control of politically sensitive content – there is common ground and evidence of consensus from Track 2 processes such as the International Dialogues on AI Safety, for the two governments to discuss this topic further.

The governments could start to agree on the domains in which evaluations are important, eventually moving towards defining common critical thresholds, such as red lines, which if crossed would signal that AI models pose significant domestic and global national security risks.

A subgroup of this dialogue could focus on the risk of advanced AI proliferation to non-state actors, and aim to build consensus on how to limit such proliferation. Both US and Chinese sources stress the importance of preventing model weight theft and a common set of actors that both countries would be concerned about are non-state actors with significant cyber capabilities.

Next, we also recommend that a series of technical standards-setting discussions be conducted, either through existing standards bodies (e.g. ISO) or new fora. There is keen attention on the need for reliability, robustness and adversarial testing from both the US and China. The scope of these discussions should be restricted to issues unrelated to national security considerations of AI systems, instead focusing largely on commercial product safety. These topics may prove especially salient as AI companies from both the US and China increasingly sell products internationally.

[140] By dangerous capabilities here we mean a cluster of AI system capabilities that may have national security implications, such as CBRN, model autonomy, cyberattacks, and persuasion.

We also recommend that existing industry coordination approaches such as the Coalition for Content Provenance and Authenticity (C2PA) consider either involving Chinese companies in existing dialogues or setting up distinct tracks of dialogue with Chinese actors. There is common concern about the need for better information traceability and watermarking across both US and Chinese governments and companies. C2PA is an existing industry association that already brings together leading companies in the US to tackle this issue, and could consider working with Chinese companies to build global norms and standards.

Finally, we recommend that Track II dialogues on emerging or novel approaches to safety and governance where there may be some existing common ground.

Dialogues focused on governance, such as the Yale Law School and Chinese Academy of Social Science Law Centre's Track II dialogue, could include discussion on innovative governance methods for AI, as part of their existing dialogues. Such dialogues could include more detailed discussion on items such as compute thresholds, and model registration systems, exchanging best practices and lessons from implementation.

More scientific dialogues such as the International Dialogues on AI Safety, could focus on the use of AI for pro-safety purposes. Our analysis suggests that there is at least some common ground on the use of AI for AI safety, such as by aiding in or even automating evaluations of new models. Scientific Track 2 dialogues, such as International Dialogues on AI Safety could leverage this common ground by discussing how AI systems can be used to improve AI safety.

5 Limitations

This study faces several challenges inherent to the rapidly evolving field of AI governance. Several policy documents we analyzed during our study have since been updated. For example, the TC260 Committee's generative AI safety standard has been updated from a finalized technical document to a draft national standard in China, with some notable changes in text.

Moreover, our focus on document substance over implementation practices, while allowing for comprehensive textual analysis, may not fully capture some real-world nuance. Moreover, unpublished contextual information about US–China perceptions heavily influences policy formulation and dialogue. Our analysis is very much an 'outside-in' study, which does not account for the attitudes of diplomats and Track II practitioners who are conducting the dialogue.

Importantly, the topics we have identified on which there is common ground are not necessarily the same as topics on which further discussion would be valuable. Dialogue could be tractable in some areas where there is currently limited common ground, simply because both countries believe there is significant cause for concern. On the other hand, dialogue could be unnecessary in some areas of agreement, for example international cooperation on user transparency may not really be productive, despite both sides being concerned about risks from limited user transparency domestically. While we do identify specific areas that are promising for dialogue, the simple fact of agreement does not by itself imply that dialogue will be advantageous. Nor does it imply an endorsement of the governance approaches discussed - further work needs to examine the merit and limitations of these approaches.

Our scope is also primarily limited to high-level domestic policies issued by central / federal governments and whitepapers or preprints issued by AI companies, all of which must specifically be about AI. This potentially overlooks convergences at lower governmental levels or ideas expressed in other types of documents issued by companies (e.g., blogposts). We also therefore exclude related legislation that is not AI-specific, such as China's Personal Information Protection Law, which is likely to bear on AI governance through its provisions on data. We also exclude intergovernmental documents and policy-oriented academic articles.

Despite these limitations, we believe our analysis offers a comprehensive foundation for understanding high-level governmental and corporate priorities in AI governance in the US and China. Given limited existing work in this area, we believe our study will be useful for researchers, as well as practitioners who are looking for grounded recommendations on topics to include in upcoming dialogues.

6 Conclusion

This paper contributes to ongoing efforts to address one of the most challenging and critical international governance issues of our time: fostering meaningful cooperation between the United States and China in managing the risks associated with advanced AI systems. While intense competition between these two countries is likely to persist, we show that such fundamental disagreements need not preclude cooperation on all fronts.

Our analysis, based on a comprehensive review of over 40 primary AI policy and corporate governance documents reveals that in certain, more narrowly-defined areas, there is common ground between the US and China, and productive dialogue may be possible. Successful dialogue on these issues could prove not only mutually beneficial but also crucial for achieving global security against AI risks[141]. By focusing on these specific issues, the US and China can make tangible progress towards effective AI governance, even in the current climate of mutual suspicion.

While our study provides a foundational framework for US–China cooperation on AI governance, it is not without its limitations. Future research should expand on this work by examining a wider range of policy documents, including more recent publications, and by conducting interviews with policymakers and Track 2 coordinators to gain deeper insights into the potential for collaboration.

In conclusion, while challenges to US–China cooperation on AI governance are substantial, our research indicates that there are promising topics for productive engagement. By prioritizing these topics for dialogue both nations can work towards mitigating global AI risks while potentially laying the groundwork for broader cooperation in the future. As AI technologies continue to advance rapidly, such dialogue may become increasingly critical for ensuring the safe and beneficial development of AI on a global scale.

[141] Yoshua Bengio et al., ”Managing Extreme AI Risks Amid Rapid Progress,” *Science* 384, 842-845(2024), 10.1126/science.adn0117

7 Acknowledgements

We are grateful for the input and feedback on earlier drafts of this report provided by the following individuals. Please note that mentioning individuals and organizations here does not imply their endorsement of this report's content.

Our thanks to Boxi Wu, Conor McGurk, Jason Zhou, Karson Elmgren, Kayla Blomquist, Oliver Guest, and Scott Singer.

Special thanks to Zachary Arnold from the Center for Security and Emerging Technology for sharing the taxonomy from the AGORA project with us and allowing us to adapt it for this project.

A The broader US–China relationship

A.1 Politics and values

The US and China exhibit fundamental divergences in their political systems and core values, contributing to bilateral tensions and potentially divergent goals for AI governance. While a comprehensive comparison of their value systems is beyond the scope of this paper, here we summarize some high-level key variables that bear on our analysis.

At a broad level, the US operates as a multi-party democratic system, emphasizing individual rights, freedom of speech, and separation of powers. China, in contrast, functions as a one-party socialist republic led by the CCP, prioritizing collective interests and social stability under the concept of "socialist democracy."[142]

The stated foreign policy aims of the US and China often appear similar, including a respect for sovereignty, human rights, and international law and a commitment to global economic development. In practice, however, the countries diverge and often come into diplomatic conflict. In particular, China often chafes at the US's perceived domination of international institutions, while the US argues that China fails to abide by international norms.[143] This tension can make it difficult to find agreement on international affairs.[144]

These divergent worldviews are apparent in the language that diplomats from the two countries use when they participate in international forums. China often advocates for a "community of shared future for mankind" and "win-win cooperation," concepts prominently featured in President Xi Jinping's speeches at forums like the UN General Assembly.[145] The US, particularly since the post-World War II era, emphasizes the importance of a "rules-based international order," a concept reaffirmed in recent national security strategies.

These differences in values and political systems underpin many of the strategic disagreements between the two nations, influencing their approaches to global governance, economic relations, and security issues, and provide the frame within which opportunities for cooperation should be analyzed.

[142] Maizland, Lindsay and Eleanor Albert. "The Chinese Communist Party," *Council on Foreign Relations Backgrounder* 6 October 2022 https://www.cfr.org/backgrounder/chinese-communist-party

[143] Gewirtz, Paul. "China, the United States, and the future of a rules-based international order," *Brookings*, 22 July 2024 https://www.brookings.edu/articles/china-the-united-states-and-the-future-of-a-rules-based-international-order/

[144] Hale, Thomas, David Held and Kevin Young. *Gridlock: Why Global Cooperation is Failing When We Need it Most* (Cambridge: Polity Press, 2013).

[145] Stella Chen, "The CMP Dictionary: Community of Common Destiny for Mankind," *China Media Project,* August 2021, https://chinamediaproject.org/the_ccp_dictionary/community-of-common-destiny-for-mankind/

A.2 Geopolitics and security

Geopolitics and security are naturally fraught areas, with key examples being maritime and nuclear security.

Maritime security issues in the Indo-Pacific region are a significant source of tension. China has extensive territorial claims across the South China Sea, which has led to disputes with several Southeast Asian nations. China asserts that under the UN Convention on the Law of the Sea (UNCLOS), military activities in exclusive economic zones (EEZs) are prohibited. In contrast, the US and many Southeast Asian nations interpret UNCLOS as allowing freedom of navigation through EEZs, including military activities, without the need to inform claimant countries. China has undertaken extensive island-building and military infrastructure construction in disputed areas of the South China Sea, particularly in the Spratly Islands, Paracel Islands, and Scarborough Shoal. Tensions also exist in the East China Sea, where China and Japan dispute the sovereignty of the Senkaku Islands (known as Diaoyu Islands in China).[146] The Taiwan Strait is another flashpoint, with increasing military activities in response to perceived moves towards Taiwanese independence. In May 2023, China conducted military drills simulating an attack on Taiwan following the inauguration of Lai Ching-te as Taiwan's president.[147]

Nuclear security is another critical area of contention. China's stance on nuclear proliferation has evolved significantly over time. During much of the Cold War, China viewed nuclear proliferation as a means to challenge superpower hegemony. However, by the 1990s, China began to cooperate more with the US on non-proliferation efforts, including attempts to limit North Korea's nuclear program in the early 2000s. As US–China relations deteriorated in the mid-2010s, official nuclear security cooperation diminished, although unofficial Track II dialogues have continued.[148] Under President Xi Jinping, China has started a massive nuclear build-up, aiming to amass 1,000 nuclear warheads by 2030, from just 200 in 2019. As nuclear security expert Tong Zhao notes, during this period President Xi has also "elevated the missile force to the status of a full military service, issued specific instructions to expedite nuclear modernization, and boosted both the sophistication and the size of China's nuclear arsenal."[149] Despite this build-up, both the US and China reaffirmed the principle that a nuclear war cannot be won and must never be fought, alongside France, Russia and the United Kingdom, in 2022.[150]

While there is fairly limited real cooperation between the US and China on security and

[146] U.S. Library of Congress, Congressional Research Service, *U.S.-China Strategic Competition in South and East China Seas: Background and Issues* by Ronald O'Rourke, R42784 (2024), https://sgp.fas.org/crs/row/R42784.pdf

[147] "What's behind China-Taiwan tensions?" *BBC*, May 2024, https://www.bbc.co.uk/news/world-asia-34729538

[148] Robert Einhorn, "Revitalizing Nonproliferation Cooperation With Russia and China," *Arms Control Association,* November 2020, https://www.armscontrol.org/act/2020-11/features/revitalizing-nonproliferation-cooperation-russia-china; "NTI Experts Participate in China-U.S. Track II Dialogue on Nuclear Security," *Nuclear Threat Initiative,* February 2024, https://www.nti.org/news/nti-experts-participate-in-china-u-s-track-ii-dialogue-on-nuclear-security/

[149] Tong Zhao, "The Real Motives for China's Nuclear Expansion," *Foreign Affairs,* May 2024, https://www.foreignaffairs.com/china/real-motives-chinas-nuclear-expansion

[150] Shannon Bugos and Julia Masterson, "NPT Nuclear-Weapon States Reject Nuclear War," *Arms Control Association,* January/February 2022, https://www.armscontrol.org/act/2022-01/news/npt-nuclear-weapon-states-reject-nuclear-war

core, likely irreconcilable differences make cooperation extremely challenging, there has been a thaw in relations from late 2023. Military-to-military talks are set to resume after two years,[151] with a key impetus being the risk of escalation in the South China Sea,[152] while on nuclear security, Chinese and US nuclear security officials held arms control talks for the first time in many years in late 2023, prior to a Biden-Xi meeting in San Francisco.[153]

A.3 Economics and trade

The US–China economic relationship has been marked by growing tensions in recent years, with an increasing number of restrictions on bilateral trade flows.

A significant number of these have been motivated by concerns around unfair Chinese trade practices hurting American industry. In 2018, the Trump administration imposed a series of tariffs on a range of goods from China, inaugurating what is often characterized as a 'trade war' between the two countries.[154] This pattern has continued through the Trump and Biden administrations, with a recent announcement in May 2024, placing tariffs on 14 Chinese goods, including renewable technologies, semiconductors and healthcare products, with a 100% tariff on Chinese electric vehicles.[155]

In addition to blocking imports, concerns about IP theft and technology transfer have led the US to place restrictions on exports to China.[156] These restrictions have come in the form of export controls and sanctions targeting the high technology sector. The Trump administration sanctioned a number of Chinese technology companies, most notably the telecommunications giant Huawei.[157] In two rounds of sanctions in 2022 and 2023, the Biden administration expanded these restrictions by implementing extensive export controls on semiconductor chips and manufacturing equipment, with the aim of preventing China from accessing and producing high end chips.[158]

This partial 'decoupling' of the US and Chinese economies is also fueled by a recognition in both countries of the dangers of interdependence in certain areas.[159] In the US, supply

[151] "US, China to resume military-to-military talks in 'coming months': Austin," *Al Jazeera*, May 2024, https://www.aljazeera.com/news/2024/5/31/taiwan-south-china-sea-dominate-meeting-of-US--China-defence-chiefs

[152] Laura Bicker, "South China Sea Tensions Force US and Beijing to Talk More," *BBC*, June 2024, https://www.bbc.co.uk/news/articles/cqvvxzv24pqo

[153] Rajeswari Pillai Rajagopalan, "China-US Nuclear Arms Control Talks: A Much-Needed First Step," *The Diplomat,* November 2023, https://thediplomat.com/2023/11/china-us-nuclear-arms-control-talks-a-much-needed-first-step/

[154] Chad P. Bown, "Four years into the trade war, are the US and China decoupling?" *Peterson Institute for International Economics,* October 2022, https://www.piie.com/blogs/realtime-economics/four-years-trade-war-are-us-and-china-decoupling

[155] "US–China Relations in the Biden Era: A Timeline," *China Briefing,* July 2024, https://www.china-briefing.com/news/US--China-relations-in-the-biden-era-a-timeline/

[156] Alan O. Sykes, "The Law and Economics of "Forced" Technology Transfer and Its Implications for Trade and Investment Policy (and the U.S.–China Trade War)," *Journal of Legal Analysis* 13 no. 1 (2021): 127-171, https://law.stanford.edu/publications/the-economics-of-forced-technology-transfer-ftt-and-its-implications-for-trade-and-investment-policy-and-the-u-s-china-trade-war/

[157] Jeremy Ney, "United States Entity List: Limits on American Exports," *Harvard Kennedy School,* February 2021, https://www.belfercenter.org/publication/united-states-entity-list-limits-american-exports

[158] Emily Benson, "Updated October 7 Semiconductor Export Controls," *Center for Strategic and International Studies*, October 2023, https://www.csis.org/analysis/updated-october-7-semiconductor-export-controls

[159] Jon Bateman, "U.S.-China Technological "Decoupling": A Strategy and Policy Framework,"

chain disruptions during the Covid-19 pandemic created additional pressure to onshore or 'friendshore' critical supply chains to the territory of the US and its allies.[160] In China, the dependence of the economy on foreign suppliers for cutting edge semiconductor chips, made apparent in the wake of US export restrictions, has led to greater investment in building a domestic industry.[161]

Despite these tensions, however, decoupling remains partial, as both economies remain tightly interlinked in many areas. US imports from China recovered somewhat after the initial tariffs, and the dollar value of exports has risen slightly.[162] The Biden administration insists that its strategy is not wholesale decoupling but rather a "small yard, high fence" approach, using targeted controls to pursue specific objectives in maintaining US technological leadership.[163] While the possibility of harder decoupling in the future cannot be discounted, at present the two economies remain closely connected.

A.4 Climate and society

The US–China climate relationship, pivotal given their combined 40% share of global CO2 emissions, is characterized by a tension between shared climate imperatives and divergent economic priorities.[164] This core conflict manifests in their contrasting approaches to fossil fuels: China's continued coal expansion versus US pressure for rapid phase-out, evident in their COP28 disagreement over "phase-out" versus "phase-down" language. Yet, this tension coexists with significant cooperation, exemplified by their joint 2014 announcement that catalyzed the Paris Agreement and their 2023 Sunnylands commitment to triple renewable capacity by 2030.[165]

Key enablers for this cooperation include sustained interactions at the highest levels between Chinese and American officials, with US Special Envoy for Climate John Kerry attributing the success of the blueprint for climate cooperation in the Glasgow Declaration in part to the 30 meetings between him and China's climate envoy Xie Zhenhua leading up to the agreement.[166] While climate goals often clash with domestic economic concerns, both nations

Carnegie Endowment for International Peace, April 2022, https://carnegieendowment.org/research/2022/04/US--China-technological-decoupling-a-strategy-and-policy-framework?lang=en

[160] Niels Graham and Mondrita Rashid, "Is 'Friendshoring' Really Working?" *Atlantic Council,* July 2023, https://www.atlanticcouncil.org/blogs/new-atlanticist/is-friendshoring-really-working/

[161] Daniel Araya, "Will China Dominate the Global Semiconductor Market?" *Centre for International Governance Innovation,* January 2024, https://www.cigionline.org/articles/will-china-dominate-the-global-semiconductor-market/

[162] Anshu Siripurapu and Noah Berman, "The Contentious U.S.-China Trade Relationship," *Council on Foreign Relations,* May 2024, https://www.cfr.org/backgrounder/contentious-US--China-trade-relationship

[163] Jake Sullivan, "Remarks by National Security Advisor Jake Sullivan on Renewing American Economic Leadership at the Brookings Institution," *The White House*, April 2023, https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/04/27/remarks-by-national-security-advisor-jake-sullivan-on-renewing-american-economic-leadership-at-the-brookings-institution/

[164] Thom Woodroofe and Brendan Guy, "Climate Diplomacy under a New U.S. Administration," *Asia Society Policy Institute,* April 2020, https://asiasociety.org/policy-institute/climate-diplomacy-under-new-us-administration-0

[165] Sara Schonhardt and Zack Colman, "They're Talking, But a Climate Divide Between Beijing and Washington Remains," *Politico*, November 2023, https://www.politico.eu/article/cop28-climate-divide-beijing-china-us-washington-john-kerry-xie-zhenhua-joe-biden-xi-jinping/

[166] Lisa Friedman, "What Happened at COP26 on Wednesday: China and U.S. Say They'll 'Enhance' Climate Ambition,"

have consistently returned to climate cooperation as a rare arena for positive engagement, establishing both high-level working groups and subnational communication channels.[167]

The societal dimensions of US–China relations have been harder to make progress on. Pew Research Center data shows unfavorable views of China among adult Americans reached 81% in 2024, while Chinese sentiment towards the US has similarly soured.[168] In terms of educational exchanges, between 2018 and 2023, the number of Chinese students in the U.S. decreased by 83,000 - over 22% while in the 2021-22 school year, there were only 211 American students in mainland China.[169] There have also been deep challenges in terms of cooperation on stemming the flow of fentanyl and opioid precursors from China to the US. While China agreed to stem the flow of fentanyl at the start of the 2020s, the US has argued that China has refused to mount adequate internal enforcement and cooperate with the US over the past two years. The Sunnyland Summit in November 2023, marked a breakthrough in cooperation on the issue, with the US–China counternarcotics commission holding its first first meeting in January 2024.[170]

B Methodology

This appendix outlines the methodology used for the results section.

Identification and selection of relevant documents

The documents that constituted our data set came from both government and corporate entities. We collected a set of government documents from two main sources - the Digital Policy Alert database and the appendices of Concordia AI's 2023 State of AI Safety in China report. These were regulatory documents from a variety of agencies and institutions within the US and China's respective government ecosystems. We also conducted a separate search for corporate governance documents (e.g., whitepapers) and model release papers that shed light on how companies think about AI governance and safety.

New York Times, November 2021, https://www.nytimes.com/live/2021/11/10/world/cop26-glasgow-climate-summit

[167] Shi Jiangtao, "US–China Joint Climate Action Back on Track, Hours Ahead of Xi-Biden Meeting in San Francisco," *South China Morning Post,* November 2023, https://www.scmp.com/news/china/diplomacy/article/3241674/US--China-joint-climate-action-back-track-hours-ahead-xi-biden-meeting-san-francisco; "China, U.S. to Establish Communication Channel For Subnational Climate Cooperation," *China Global Television Network,* June 2024, https://news.cgtn.com/news/2024-06-18/China-U-S-to-establish-subnational-climate-cooperation-channel-1uxfMatecik/p.html

[168] Christine Huang, Laura Silver, and Laura Clancy, "Americans Remain Critical of China," *Pew Research Center,* May 2024, https://www.pewresearch.org/global/2024/05/01/americans-remain-critical-of-china/

[169] Brian Wong, "Why Sino-American Education Exchanges Matter, and How to Revive Them," *China-United States Exchange Foundation,* March 2024, https://www.chinausfocus.com/society-culture/why-sino-american-education-exchanges-matter

[170] Vanda Felbab-Brown, "US–China relations and fentanyl and precursor cooperation in 2024," *Brookings Institution,* February 2024, https://www.brookings.edu/articles/US--China-relations-and-fentanyl-and-precursor-cooperation-in-2024/

Inclusion and exclusion criteria

We prioritized the collection of recent documents where possible (i.e. released in 2023 or after) to capture the latest thinking on generative AI. In cases where there was sparser data, we cast further back to 2021. This was especially the case with corporate policy or governance documents in China, where there was much less contemporary data available. In certain cases, where focusing solely on documents released during or after 2023 would leave out key documents (e.g., China's recommendation algorithm regulations), we also made specific exceptions to include them.

For government documents we only included official, national-level documents, excluding national-level draft legislation or legislative proposals. For example, we excluded numerous local government documents on AI policy from cities such as Beijing, and states such as California. While developments in these sub-national jurisdictions may be influential domestically, we focused our study on documents that were likely to provide the most signal on what the US and Chinese governments could cooperate on at an international level. We did include two AI law proposals from China, which appeared to receive significant attention domestically and could provide some signal on where national level AI policy might be heading.

On the corporate side we included select pieces of evidence. Broadly, we only included model release papers issued by leading frontier AI companies in both China and the US as these were likely to contain the most up-to-date description of the company's approach to safety. In addition to this, we also included select AI governance whitepapers from Chinese companies to provide further clarity on the Chinese corporate interpretation of Chinese legislation and the Chinese corporate approach to governance and safety.

Finally, we also excluded international statements or declarations, from our dataset, given that these statements tend to be curated for an international audience. We focus instead on looking through documents primarily intended for domestic audiences, which are likely to provide much more detailed and substantive understanding of AI policy concerns.

Other sources we excluded were think-tank reports and judicial decisions or case law.

Coding process

Our coding process adapted a taxonomy for categorizing actors' AI risk and governance approaches from CSET's AGORA project. This taxonomy listed out several different key risks from AI (e.g., bias, risks from AI-generated content), and governance approaches (e.g., convening), including definitions and examples. We adapted this by adding in a few new categories and specifying a few definitions, as our coding was at the 'span-level' (i.e. tagging specific quotes) not at the 'document-level'.

Each of the researchers on this project read documents line-by-line, tagging spans according to the relevant taxonomy (risks and/or governance approaches). The same quotes could be tagged for both risks and governance approaches, and if a quote involved more than one risk or governance approach, that would be entered twice in the dataset.

Researchers flagged uncertain quotes for review by other researchers in the team, and sample checks were also performed by an editor across all documents, in order to harmonize the coding standards across the different researchers.

We did not employ detailed annotation or coding standards beyond using standardized definitions of categories and illustrative examples to align the approaches used by the different researchers on the team. As such there is a degree of subjectivity in the coding process, as researchers relied on their subjective judgment to decide which quotes should be included in the dataset.

Dataset Description

Below, we provide a breakdown of the dataset by type of issuing entity, showing the number of documents, total number of quotes and average number of quotes per document.

Table 6: Breakdown of the documents in the dataset by type of issuing entity

Country of issuance	**Corporate**	**Executive (Party)**	**Executive (State)**	**Legislature**	**Others**	**Grand Total**
China	13	3	11		3	**30**
USA	3		9	1	1	**14**
Grand Total	**16**	**3**	**20**	**1**	**4**	**44**

Table 7: Breakdown of the quotes in the dataset by type of issuing entity

Country of issuance	**Corporate**	**Executive (Party)**	**Executive (State)**	**Legislature**	**Others**	**Grand Total**
China	115	25	57		44	**241**
USA	22		161	2	4	**189**
Grand Total	**137**	**25**	**218**	**2**	**48**	**430**

Our dataset consisted of about double the number of Chinese documents as US documents, accounted for by the larger number of corporate documents and slightly higher number of executive documents from China. Our US sample contained the latest available model cards from the three leading frontier AI companies (OpenAI, Anthropic and Google Deepmind), whereas for China, given a less clear sense of the frontrunners in the AI race, the dataset

Table 8: Breakdown of the average number of quotes per document in the dataset by type of issuing entity

Country of issuance	**Corporate**	**Executive (Party)**	**Executive (State)**	**Legislature**	**Others**	**Grand Total**
China	9	8	5		15	**8**
USA	7		18	2	4	**14**
Grand Total	**9**	**8**	**11**	**2**	**12**	**10**

contains a wider range of model release papers and corporate whitepapers. There are slightly more Chinese state and party documents because significant elements of Chinese AI governance today are continuous with and evolve from pre-2023 regulations (e.g., the algorithm registry).

Despite the larger number of Chinese documents, there is only a small difference in the number of quotes extracted. The largest differences are in the significantly higher number of quotes from Chinese corporate documents, explained by the inclusion of corporate whitepapers from Chinese AI labs, and the significantly higher number of US executive branch quotes, explained by extensive reference to the Biden AI EO. We drew heavily on these sources because they provided the best source of information about these respective actors in each context.

Finally the average number of quotes extracted per document is heavily influenced by the length of the documents consulted, and more specifically, the length of the relevant portions of these documents. The reason 'Others' in China has a significantly higher number of quotes on average is explained by the inclusion of two long draft AI laws written by experts in China, which provide some signal on the direction of Chinese AI regulation. In the US dataset, the high average number of quotes per document for the executive branch is explained by the inclusion of the long Biden AI EO as part of the dataset.

Analysis

After the entire dataset was coded, researchers manually compared Chinese and US quotes on AI risk and governance strategies, using the criteria in the table below to grade the degree of overlap between US and Chinese sources. In our findings, the degree of overlap between US and Chinese actors for each risk and governance strategy is justified through a description of the data.

Table 9: Criteria for assessing degrees of overlap between US and Chinese documents

Degree of overlap	**Criteria**
Strong	Most, if not all, of the material found in the dataset shows concurrence between US and Chinese understanding of a given risk or governance approach
Moderate	While there is some overlap in understanding, there are also key differences that we could identify in the way concepts were raised
Weak	Majority of the evidence pointed towards differing understandings of emphases on a given issue