IAPS | Institute for AI Policy and Strategy

May 2nd – 2024

Topics for track IIs

What can be discussed in dialogues about advanced AI risks without leaking sensitive information?

AUTHORS

Oliver Guest – Research Analyst

Zoe Williams – Acting Co-Director

Executive Summary

There is significant concern in both the U.S. and China that AI systems may pose catastrophic risks from accidents or misuse. U.S.-China "track IIs" about managing risks of advanced AI might be a promising way to reduce these risks.[1]

Possible downsides of such track IIs include that they might contribute to proliferation and unintended capabilities disclosure. We use "proliferation" to mean spreading information about how to build AI systems with increased offensive or dual-use capabilities. We use "unintended capabilities disclosure" to mean the unintended spread of information about the level of a country or developer's offensive or dual-use AI capabilities. This disclosure might be disadvantageous from a national security standpoint.

In this issue brief, we highlight topics that would be valuable to discuss in track IIs and that are unlikely to contribute to proliferation or unintended capabilities disclosure. Focusing on these topics could make it possible to have valuable track IIs, even if avoiding proliferation or unintended capabilities disclosure is a priority.

The highlighted topics are summarized in the table below.

Theme	Topic
Developer-level AI governance	• Sharing lessons learned for implementing specific developer best practices (e.g., pre-deployment risk assessment, capabilities evaluations, third-party model audits, and red teaming).[2] • Surfacing new best practices. • Discussing overarching governance frameworks (e.g., "Responsible Scaling Policies").
National-level AI governance	• Sharing best practices and lessons learned from AI regulation efforts in specific countries. • Identifying which AI risks are best dealt with at the national level.
International-level AI governance	• Identifying which AI risks are best tackled at the international level. • Discussing what institutions or international agreements need to be created or adapted to better reduce AI risks.

[1] Track II diplomacy refers to non-governmental, informal, and unofficial contacts and activities between private citizens or groups of individuals.
[2] These examples are from Schuett et al., "Towards Best Practices in AGI Safety and Governance."

	• Establishing consensus on "red lines" for AI development and deployment that should not be crossed.
Non-proliferation measures	• Sharing best practices relating to some non-proliferation measures (e.g., publication norms). • Exploring possible joint efforts to prevent AI proliferation (e.g., institutions analogous to the IAEA or NSG).[3] • Discussing how to avoid harmful proliferation to third countries while still sharing the benefits of AI globally.
Detail about safety risks	• Discussing potential pathways to catastrophic AI accidents or misuse. • Examining the implications of recent technical research about AI risks.
Technical safety methods	Sharing technical insights about safety, focusing specifically on safety approaches that have low "capabilities externalities," such as: • Multi-agent safety techniques to ensure AI systems remain safe when interacting. • Power aversion methods to prevent AI systems from seeking excessive influence. • Anomaly detection to identify when AI systems are behaving in potentially hazardous ways.
Model safety evaluations	• Discussing governance frameworks for rigorous model safety evaluations. • Sharing technical details about carrying out evaluations for capabilities that could help misaligned AI systems evade oversight.

[3] The International Atomic Energy Agency and Nuclear Suppliers Group both play a role in nuclear non-proliferation efforts.

Introduction

There is significant concern in both the U.S. and China that AI systems may pose catastrophic risks from accidents or misuse. In both countries, concerns have been raised at the highest levels of government,[4] by academics,[5] and—particularly in the U.S.—by leaders within the AI industry.[6]

"Track IIs" have been proposed as one way to reduce these risks. Track II diplomacy refers to non-governmental, informal, and unofficial contacts and activities between private citizens or groups of individuals.[7] This issue brief focuses on track IIs that aim to manage risks of advanced AI. For concision, we generally refer to these just as "track IIs." We focus, in particular, on track IIs that include participants from the U.S. and China.[8] That said, we expect that much of our analysis would also apply to dialogues between different groupings of countries.

Track IIs might reduce advanced AI risks in various ways. For example, they could be helpful for identifying and sharing best practices that AI developers or other actors should take to reduce these risks. Track IIs could also be helpful stepping stones towards institutions that might help reduce these risks, such as an "IPCC for AI."

However, there are also plausible ways in which track IIs might be harmful. We focus here on two that we expect might be particularly concerning to policymakers:

- **Proliferation**: Track IIs might proliferate information about how to build AI systems with higher offensive or dual-use capabilities, such as the ability to write sophisticated

[4] "President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence"; "The Bletchley Declaration"; Heide, "Beijing Policy Interest in General Artificial Intelligence Is Growing."
[5] "Prominent AI Scientists from China and the West Propose Joint Strategy to Mitigate Risks from AI"; Bengio et al., "Managing AI Risks in an Era of Rapid Progress."
[6] Senior figures in several U.S. AI companies have repeatedly said that AI might pose catastrophic risks, such as in the CAIS statement. We are aware of fewer examples from leaders of Chinese AI companies, though there were some corporate signatories of the IDAIS-Beijing statement, and Concordia AI lists several other examples. "Statement on AI Risk"; Safe AI Forum, "International Dialogues on AI Safety"; Concordia AI, "State of AI Safety in China," 53–57.
[7] Jones, *Track Two Diplomacy in Theory and Practice*, 9. Similar to track IIs are track 1.5s; these are unofficial but involve some government officials. We expect that many of the points in this issue brief would also apply to track 1.5.
[8] Dialogues between participants from these countries might be particularly important for two reasons. First, they are among the leaders in AI development, giving them special responsibility for mitigating potential risks. Second, they have a strained relationship, meaning that it may take more effort to achieve any cooperation or information-sharing that is needed between them to reduce risks. Guest, Aird, and Heide, "International AI Safety Dialogues: Benefits, Risks, and Best Practices," 10.

compute code, including for cyber attacks.[9] Dual-use foundation models are a key example in this category.

- **Unintended capabilities disclosure**: Governments may have national security reasons for keeping confidential information about their country's level of offensive and dual-use AI capabilities.[10] However, track IIs could contribute to the disclosure of this information against the government's intent, potentially undermining these security objectives.

Various commentators, in both the West and in China, have suggested topics that might be productive to discuss at track IIs about managing advanced AI risks.[11] Topics have also been suggested for the U.S.-China track I on AI.[12] In our issue brief, we aim to specifically identify topics that are unlikely to contribute to proliferation or unintended capabilities disclosure. There are many topics that can be productively discussed, even if avoiding these potential downsides is a priority for decision-makers.[13]

Scope of this issue brief:

- We focus on proliferation or unintended capabilities disclosure that might happen as a result of topics on the agenda at track IIs. This excludes, for example, proliferation that happens because attending a track II makes participants more vulnerable to espionage.[14]

---

[9] By "dual-use", we mean AI systems that can be put toward both civilian and military uses, and more broadly, toward beneficial and harmful ends. Brundage et al., "The Malicious Use of Artificial Intelligence," 16.

[10] As discussed below, increased disclosure about different actors' level of capabilities might also contribute to more intense racing dynamics to develop AI systems, though this is less clear.

[11] Concordia AI, "The State of China-Western Track 1.5 and 2 Dialogues on AI"; Guest, Aird, and Heide, "International AI Safety Dialogues: Benefits, Risks, and Best Practices," 31–35; Imbrie and Kania, "AI Safety, Security, and Stability Among Great Powers," 7–9; Kissinger and Allison, "The Path to AI Arms Control"; Ying and Allen, "Together, The U.S. And China Can Reduce The Risks From AI." There are also some descriptions of topics that have been discussed in previous dialogues, such as the Brookings-CISS Tsinghua dialogue. Center For International Security And Strategy, Tsinghua University, "The China-U.S. Track II Dialogue on Artificial Intelligence and International Security Interim Report."

[12] Wang and Zhu, "What Topics Can Be Discussed in the China-U.S. Artificial Intelligence Dialogue [中美人工智能对话可以谈些什么]"; Hass and Kahl, "Laying the Groundwork for U.S.-China AI Dialogue"; Webster and Hass, "A Roadmap for a U.S.-China AI Dialogue."

[13] Proliferation and unintended capabilities disclosure are not the only reasons why readers might be concerned about track IIs. For example, U.S. policymakers sometimes complain the U.S. has to make concessions to China in order to even have dialogues. As a result, some readers may be skeptical of track IIs about managing advanced AI risks, even if they accept all our arguments here. "Select Committee Republicans Issue Demands For Xi Jinping Ahead of Meeting with President Biden"; Schneider and Hauser, "Pottinger on Trump 2.0."

[14] We would be pleased to see analysis of how important these other mechanisms are and what could be done to address them.

- There may be trade-offs between avoiding proliferation or unintended capabilities disclosure and achieving particular benefits, such as reduced accident risk.[15] We leave out of scope the question of how much avoiding proliferation and unintended capabilities disclosure *should* be prioritized over possible benefits.

In the rest of this issue brief, we first expand on the possible concerns about proliferation and unintended capabilities disclosure. We then highlight topics that could be usefully discussed in track IIs and that are unlikely to contribute to either of these potential issues.

[15] For example, there are some technical safety techniques that make AI systems more useful. Discussing these techniques in detail might reduce AI accident risks but at the cost of higher proliferation risk.

Expanding on possible concerns

Proliferation

One risk of track IIs is that they might contribute to the proliferation of offensive or dual-use AI capabilities to geopolitical rivals. We use "dual-use" to mean AI systems that can be put toward both civilian and military uses and, more broadly, toward beneficial and harmful ends. An example of an offensive AI capability might be the ability to carry out cyber attacks. An example of a dual-use AI capability might be the ability to write sophisticated computer code—including for cyber attacks.

Dual-use foundation models

A key example of a dual-use AI system is a "dual-use foundation model."[16] The 2023 AI Executive Order defines this kind of model and imposes various obligations to reduce the risks from this kind of model.

The definition includes that the model is "applicable across a wide range of contexts" and could exhibit "high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety." Dangerous tasks could include lowering the barrier of entry for acquiring CBRN weapons or enabling powerful offensive cyber operations.

It is likely that successors to existing large language models (such as GPT-4) would qualify as dual-use foundation models.[17]

As an example of how proliferation could occur, some safety techniques (such as reinforcement learning from human feedback) improve our ability to steer AI behavior; this reduces safety risks from AI systems but can also make them straightforwardly more useful to developers or end users.[18] Transferring such technical insights to geopolitical rivals would not just reduce safety

[16] Other examples include that drug-discovery models can be used both to find medicines and to propose new biochemical weapons. Urbina et al., "Dual Use of Artificial-Intelligence-Powered Drug Discovery."
[17] Existing large models can already be used in a wide range of contexts. AI developers have stated that, in the absence of risk mitigations, their future models may pose major security risks in several ways, including via making it easier to produce CBRN weapons. OpenAI, "Preparedness"; Anthropic, "Anthropic's Responsible Scaling Policy, Version 1.0."
[18] Christiano et al., "Deep Reinforcement Learning from Human Preferences"; Christiano, "Thoughts on the Impact of RLHF Research."

risks but also help them build AI systems that are more useful to them, including offensive or dual-use systems.[19]

There are national security reasons that might motivate concern about the proliferation of offensive or dual-use AI systems to geopolitical rivals. One might be concerned that these systems would be weaponized or that they would increase the competitiveness of rivals more broadly, such as by helping them with economic competition. Proliferation might also increase the likelihood of catastrophic AI accidents.[20] For example, it might increase racing dynamics around AI development by narrowing the competition between AI developers. More intense racing dynamics might incentivize cutting corners on safety.[21] Additionally, if there are more actors with the knowledge to develop powerful AI systems, there might be a higher likelihood that one actor is careless enough that their AI system contributes to an accidental catastrophe.[22]

Unintended capabilities disclosure

Track IIs may also contribute to unintended capabilities disclosure. We define this as information being disclosed, contrary to a government's intention, about the level of offensive or dual-use AI capabilities available to that country.[23]

Governments may sometimes *want* information about these capabilities to be disclosed.[24] For example, governments might deliberately share such information with each other as part of confidence-building measures or arms control agreements for AI.[25] There are various historical

[19] Some safety insights could also be helpful for increasing capabilities by making the AI system fundamentally more capable, not just more steerable by developers and/or users. For example, safety-motivated work on interpretability seems to have contributed to advances on increasing context length. Poli et al., "Hyena Hierarchy"; Räuker et al., "Toward Transparent AI," 11.
[20] "Avoiding Extreme Global Vulnerability as a Core AI Governance Problem." AI accidents might have transnational effects; this presents a compelling reason for governments to be concerned about accidents outside their jurisdiction, not just within it. Trager et al., "International Governance of Civilian AI: A Jurisdictional Certification Approach," 11.
[21] If a developer spends resources on reducing the risks from its AI systems, these are resources that it cannot spend on making its AI systems more capable. Additionally, applying safety techniques to existing AI systems might make these systems less capable in some ways. Gleave, "AI Safety in a World of Vulnerable Machine Learning Systems"; Leike, "Distinguishing Three Alignment Taxes."
[22] Accident risk concerns do not just apply to proliferation to geopolitical rivals. However, proliferation to rivals might be particularly concerning from an accident perspective. For example, coordination not to cut corners on safety might be particularly difficult between rivals.
[23] Although we focus on governments, similar arguments might apply to AI developers. For example, AI developers might be wary of their employees attending track II events out of a concern that the employees would reveal sensitive information about what the developer has achieved internally.
[24] One could also imagine cases where governments are indifferent about such information being disclosed. Disclosure in these cases would not count as unintended.
[25] For example, the proposals in the following papers might involve sharing information about levels of dual-use capabilities: Horowitz and Scharre, "AI and International Stability"; Lamberth and Scharre, "Arms Control for Artificial Intelligence"; Shoker et al., "Confidence-Building Measures for Artificial Intelligence."

cases where the deliberate disclosure of strategically relevant capabilities had a stabilizing effect on international relations.[26] When we refer to unintended capabilities disclosure, we are not referring to information that governments choose to release.

Two mechanisms seem most plausible for how track IIs should contribute to unintended capabilities disclosure. First, individual participants might deliberately disclose this information to each other, e.g., out of a belief that it will build trust.[27] Second, it may be difficult for participants to discuss some topics at track IIs without incidentally divulging information about capabilities levels. For example, participants might want to talk about what research areas countries or companies are investing in. It might be difficult to do so without revealing some information about the various capabilities available to these actors.

From a national security perspective, unintended capabilities disclosure might be harmful in several ways. First, it could provide information that rivals could use to challenge the original country. For example, if one country's superiority in AI-enabled cyber defense is well-known, rivals might invest more heavily in sophisticated cyber offense techniques or know to devise strategies to compete that do not rely on cyber. Second, if one country reveals a particular AI-related vulnerability, such as that AI systems involved in critical infrastructure are not robust, rivals could know to target these weaknesses. Third, disclosure of military-relevant AI capabilities might reduce rivals' uncertainty about what capabilities a country has, potentially encouraging rivals to take bolder hostile actions.

Unintended capabilities disclosure *may* also contribute to racing dynamics in AI development. Increased racing dynamics might increase the likelihood of catastrophic AI accidents, such as by motivating corner-cutting on safety. That said, the effects of greater unintended capabilities disclosure on racing dynamics seem unclear. Formal models have come to different results and, in any case, make strong modeling assumptions that might not hold in reality.[28] Racing dynamics might be particularly intense if actors overestimate their rivals' level of dual-use capabilities. However, it is unclear whether increased unintended capabilities disclosure

---

[26] For examples relating to confidence-building measures, see Shoker et al., "Confidence-Building Measures for Artificial Intelligence," 6–14. Arms control sometimes has a stabilizing effect, such as via reducing the fear of a surprise attack. Arms control agreements seem to be easier to make in cases where countries are able and willing to share information about the relevant military capabilities with each other. Jervis, "Arms Control, Stability, and Causes of War," especially pp. 170-173; Coe and Vaynman, "Why Arms Control Is So Rare," especially pp. 342-343.

[27] In this case, the disclosure is intended by the participant, even if it is not what the affected government intends.

[28] See in particular Armstrong, Bostrom, and Shulman, "Racing to the Precipice." and Emery-Xu, Park, and Trager, "Uncertainty, Information, and Risk in International Technology Races." The former paper finds that increased disclosure is more dangerous. The latter (and more recent) paper finds that increased disclosure is more dangerous only if "decisiveness" is high, referring to a situation where small performance leads produce larger probabilities of being the first to develop a new technology. An example of a strong assumption in these models is the assumption that AI developers can be modeled as coherent actors that maximize expected utility.

(including from track IIs) would make overestimates less likely. We discuss these questions in the Appendix. Given these uncertainties, national security concerns seem like a significantly more robust reason than racing dynamics to be worried about unintended capabilities disclosure from track IIs.

Lower risk topics

There are many topics that could be discussed in track IIs that have a low likelihood of contributing to proliferation or unintended capabilities disclosure. We group them here by theme, such as developer-level AI governance and national-level AI governance.

Developer-level AI governance

The developers of AI systems have an important role to play in mitigating risks. This theme focuses on identifying governance practices and frameworks that they can implement to reduce risks.

Discussing developer best practices is unlikely to contribute to proliferation insofar as it is difficult to discuss the best practices without also discussing non-public technical details of technical AI systems. These discussions might *reduce* overall proliferation; various best practices that could be discussed would likely reduce proliferation if implemented.[29] Similarly, discussing best practices will have a low likelihood of unintended capabilities disclosure, apart from insofar as these discussions involve discussing non-public information about actors' AI capabilities.

Lower risk topics about developer-level AI governance

Lessons learned for implementing specific best practices. There are already many best practices for reducing AI risks that experts support with a high degree of consensus. Track IIs could be a helpful venue in which to share lessons learned about how to implement them. The following best practices had particularly high support in a 2023 expert survey by Schuett et al.:[30]

- **Pre-deployment risk assessment** to identify, analyze, and evaluate risks from powerful models before deploying them.
- **Dangerous capabilities evaluations** to assess models' dangerous capabilities (e.g., misuse potential, ability to manipulate, and power-seeking behavior).
- **Third-party model audits** before deploying powerful models.
- **Red teaming** by external teams before deploying powerful models.

[29] Examples in an expert survey include "protection against espionage," "security standards," and "no unsafe open-sourcing." Schuett et al., "Towards Best Practices in AGI Safety and Governance," 18–20.
[30] Schuett et al., "Towards Best Practices in AGI Safety and Governance."

- **Safety restrictions** for powerful models after deployment (e.g., restrictions on who can use the model, how they can use the model, and whether the model can access the internet).

Surfacing new best practices. There may be best practices that some participants do not yet know or whose value is not understood by all participants. Track IIs could be a helpful opportunity for participants to explain particular best practices to their counterparts and why they might be valuable. As an example, the expert survey cited above seems only to have involved experts from Western institutions.[31] There might be valuable best practices that Chinese experts have identified but that are not yet widely known among other experts.

Discussing overarching self-governance frameworks. As well as discussing specific best practices, track IIs could be helpful for discussing overarching frameworks that include various best practices. For example, some AI labs have committed themselves to frameworks for what risk-reduction measures they will implement at given levels of AI capabilities. Key examples include Anthropic's "Responsible Scaling Policy" and OpenAI's "Preparedness Framework."[32] Participants could discuss the value of overarching frameworks like these and how best to implement them.

Overall, discussing best practices for AI developers seems low-risk from the perspective of proliferation and unintended capabilities disclosure. However, there are some potential ways in which these discussions could contribute to these risks, particularly if participants share non-public information or go into significant technical detail.

First, if participants reveal that a developer has implemented a specific best practice, other participants might interpret this (rightly or wrongly) as a sign that the developer has achieved new capabilities that necessitate this practice. For instance, if it is disclosed that a particular developer has started red teaming for a particular dangerous capability, participants might take this as evidence that this developer's AI systems could plausibly exhibit this capability.

[31] The institutions of most participants are listed in Appendix A—all the institutions listed are in the West.
[32] Anthropic, "Anthropic's Responsible Scaling Policy, Version 1.0"; OpenAI, "Preparedness."

Second, a developer suddenly ramping up risk mitigation measures could be seen as implying a belief that this developer's AI systems now pose much higher risks. This change, if shared in a track II, might signal significantly increased capabilities to other participants.[33]

Finally, going into the specific details of a developer's safety practices could potentially reveal sensitive information about the technical characteristics of the developer's AI systems. For example, explaining the details of how a developer audits models for misuse potential might expose key facts about the models' architectures or training data. Proliferating this type of information might make it easier for others to replicate advanced AI systems.

National-level AI governance

This theme focuses on how individual countries can manage risks from advanced AI systems through domestic policies and regulations.

Several topics in this category are unlikely to involve discussing non-public technical information about AI systems or non-public information about actors' AI capabilities. Such topics are unlikely to contribute to proliferation or unintended capabilities disclosure.

Lower risk topics for national-level AI governance

Sharing best practices and lessons learned from AI regulation efforts. Countries can learn from each other's efforts to regulate advanced AI systems, even if the ideological objectives are very different. For example, participants could discuss whether AI regulation is best handled by one or several national-level regulators. China has primarily taken the former approach, whereas the U.S. and UK have primarily taken the latter. There might be valuable lessons learned from comparing these systems.[34]

[33] This point does not apply if the change in risk mitigation would be visible to observers anyway, even without discussion at track IIs. However, it does seem plausible that a change would not be easily visible to observers. For example, Anthropic describes several measures that the company would take if moving from "ASL-2" to "ASL-3." These include measures that might be hard for outsiders to observe such as having a high level of security around model weights, compartmentalizing certain information about AI systems, and implementing internal usage controls. Anthropic, "Anthropic's Responsible Scaling Policy, Version 1.0," 7–9.

[34] MacCarthy, "The US and Its Allies Should Engage with China on AI Law and Policy"; Sheehan, "What the U.S. Can Learn From China About Regulating AI."

What risks are best dealt with at the national level? Trager et al. give various reasons why high-stakes AI systems will require international governance.[35] Although these arguments indicate that international governance will often be needed, there may be cases where these arguments do not apply, and so where governance at the national level might be tractable and sufficient. Identifying such cases would be valuable because it might often be more feasible to implement risk-reduction measures at the national rather than international level. For example, it is less likely that there would be concerns about loss of sovereignty.[36]

There are additional topics around national-level AI governance that might be valuable to discuss, but where there might be a higher risk of proliferation or unintended capabilities disclosure:

- Participants could discuss with each other which AI risks most urgently need to be addressed by regulators. However, these discussions might touch on non-public information about dangerous capabilities that have been seen in not-yet-released AI systems.
- Exploring and clarifying government intentions around AI development could help build trust and reduce the risk of dangerous miscalculation or escalation. However, these discussions might also reveal sensitive information about a country's AI strategy and capabilities. Participants would need to carefully balance the benefits of transparency against the risks of disclosing details that could harm national security or contribute to racing dynamics.

International-level AI governance

Both the Chinese and U.S. governments acknowledge that many risks associated with AI are international, suggesting that these might be most effectively managed through international governance measures.[37] This theme focuses on governance mechanisms (e.g., treaties, institutions, codes of conduct) that could be established at the international level to manage these risks. Track II dialogues could help build consensus on the highest priority risks to tackle

[35] Trager et al., "International Governance of Civilian AI: A Jurisdictional Certification Approach," 11–14.
[36] Relatedly, participants could discuss which AI risks should be prioritized by national-level governance efforts, such as because they seem particularly concerning. However, this might reveal some information about non-public AI capabilities, i.e., contribute to unintended capabilities disclosure. For example, if some participants stress that AI-designed bioweapons should be a key focus of national-level governance, other participants might—correctly or incorrectly—assume, as a result, that capabilities have increased in this area.
[37] "The Bletchley Declaration"; Trager et al., "International Governance of Civilian AI: A Jurisdictional Certification Approach," 5–6.

internationally and what concrete steps could be taken to address them at the international level.

As in previous sections, discussion topics will have low proliferation risk if they do not involve discussing non-public technical details of AI systems. They will have low unintended capabilities disclosure risk if they do not involve non-public information about dual-use AI systems.

Lower risk topics about international-level AI governance.

What risks from AI are best tackled at the international level? Some AI risks are particularly likely to require international action, e.g., because they involve transnational effects of AI systems or because there are collective action problems that make it difficult for national governments to act unilaterally.[38]

What institutions or international agreements need to be created or adapted to better reduce AI risks? Common examples include new institutions for AI modeled on CERN, the IAEA, or the IPCC.[39]

"Red lines" for AI development and deployment. Experts from China, the U.S., and other countries have already identified some "red lines" that should not be crossed in the development and/or deployment of advanced AI systems.[40] It might be valuable for participants to identify further red lines or to establish consensus about the details of what it means to comply with existing red lines.

Discussions of these topics might be more productive if participants can cite non-public information about what concerning AI capabilities are most likely to emerge, e.g., based on experiments with unreleased AI systems. However, this might present a trade-off since revealing this information might also increase the likelihood of proliferation or unintended capabilities disclosure.

---

[38] Trager et al., "International Governance of Civilian AI: A Jurisdictional Certification Approach," 11–14. Participants could also discuss which AI risks should be prioritized by international-level governance efforts, such as because they seem particularly concerning. However, doing so might reveal some information about non-public AI capabilities, i.e., contribute to unintended capabilities disclosure. For example, if some participants stress that AI-designed bioweapons should be a key focus of international-level governance, other participants might—correctly or incorrectly—assume, as a result, that capabilities have increased in this area.

[39] Ho et al., "International Institutions for Advanced AI"; Maas and Villalobos Ruiz, "International AI Institutions."

[40] Safe AI Forum, "International Dialogues on AI Safety."

Non-proliferation measures

This theme focuses on how the U.S. and China can cooperate to prevent harmful AI proliferation to third countries while still ensuring that the benefits of AI are shared globally. These efforts might reduce the number of actors who can develop or deploy advanced AI systems in a reckless or malicious way. We are not referring here to discussions about efforts to prevent proliferation *between* the U.S. and China; the most prominent example here is likely the U.S. controls on exports of semiconductor manufacturing equipment (SME) to China.[41] We are unsure whether it would be productive for export controls towards China to be on track II agendas and leave this question out of scope.

The U.S. government is already attempting to avoid AI proliferation to third countries. For example, the SME export controls target many countries other than China.[42] Additionally, the U.S. government is attempting to prevent the theft of AI model weights, including by state actors.[43] It is unclear to what extent the Chinese government will prioritize non-proliferation to third countries; we discuss this question in the appendix.

Several topics related to non-proliferation can be discussed without referencing non-public information and thus with a low risk of proliferation or unintended capabilities disclosure. We expect that discussing these topics would overall *reduce* proliferation; even if the topics slightly increase proliferation between the U.S. and China, they will reduce proliferation to third countries.

- To have low proliferation risk, the topic should not involve non-public technical details of AI systems or information that could be used to evade non-proliferation measures.
- To have low unintended capabilities disclosure risk, the topic should not touch on non-public information about different actors' AI capabilities.

Lower risk topics about non-proliferation measures

How to implement (some) non-proliferation measures. Participants could share best practices that lower proliferation risk. They should be careful to focus on

[41] Allen, "Choking off China's Access to the Future of AI"; Dohmen and Feldgoise, "A Bigger Yard, A Higher Fence." Other U.S. measures that reduce the likelihood of proliferation to China include efforts to prevent the theft of model weights, including by state actors. Ee and O'Brien, "Putting New AI Lab Commitments in Context." One example from the Chinese side might be China's export controls on Germanium and Gallium, metals that are used in semiconductor manufacturing. Godek, "Why China's Export Controls on Germanium and Gallium May Not Be Effective."
[42] Dohmen and Feldgoise, "A Bigger Yard, A Higher Fence."
[43] Ee and O'Brien, "Putting New AI Lab Commitments in Context." Model weights are the key internal variables of an AI system that encode its capabilities and knowledge.

non-proliferation measures where detailed knowledge of the measure does not make the measure easier to bypass; otherwise, these efforts to lower proliferation to third countries might make it easier for the U.S. or China to evade existing non-proliferation measures. As an example, participants could discuss publication norms for AI research and the question of how best to manage the trade-off between sharing valuable research and proliferating potentially dangerous capabilities.[44] It is unlikely that such discussions would make it easier to evade the anti-proliferative effects of the relevant publication norms.[45]

Possible joint efforts to prevent AI proliferation. Track IIs could also be a stepping stone towards more ambitious cooperation on non-proliferation, such as with institutions analogous to the International Atomic Energy Agency (IAEA) or Nuclear Suppliers Group (NSG).[46] For example, participants could identify and increase consensus about whether it would be desirable to have such institutions and exactly what they should look like.

Discussing how to avoid harmful proliferation while still sharing AI's benefits. Countries have recognized the importance of sharing AI's benefits globally.[47] There may be difficult trade-offs between this objective and preventing the proliferation of dual-use AI capabilities.[48] Track IIs may provide a helpful forum for discussing how best to manage these trade-offs.

Some topics related to non-proliferation may be more risky to discuss. In particular, if participants share exploitable details of non-proliferation defenses, then other participants or organizations with which they are linked could use these details to get around the defenses,

[44] As some examples of what these discussions could look like, see Gupta, Lanteigne, and Heath, "Report Prepared by the Montreal AI Ethics Institute (MAIEI) on Publication Norms for Responsible AI." and Whittlestone and Ovadya, "The Tension between Openness and Prudence in AI Research."
[45] As another example, it seems likely that it would be possible to discuss best practices for structured access without thereby making it easier to evade the non-proliferation effects of structured access mechanisms. Structured access refers to mechanisms where actors can use an AI system to accomplish particular tasks but cannot access the system's source code or weights, reducing proliferation risk compared to an actor having full access. Shevlane, "Structured Access."
[46] Among other roles, the IAEA monitors state nuclear programs to ensure they are only for peaceful processes. The NSG is a group of countries that seeks to prevent nuclear proliferation by controlling the export of materials, equipment, and technology that can be used to manufacture nuclear weapons. Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," 12–15; Maas and Villalobos Ruiz, "International AI Institutions," 24–25.
[47] "The Bletchley Declaration."
[48] This can be framed as a specific example of the use-misuse tradeoff. Anderljung and Hazell, "Protecting Society from AI Misuse," 6.

contributing to proliferation. Detailed information about how model weights are secured might be an example in this category.

Detail about safety risks

This theme focuses on the details of how advanced AI systems could pose risks. For example, participants could share detailed descriptions of how AI accidents could come about or have in-depth discussions about the implications of new technical research about AI risks.

While the potential risks of AI remain a topic of ongoing debate, many influential figures in both the U.S. and China believe that AI systems could pose significant risks from misuse and/or accidents.[49] Discussing these risks in detail could be beneficial for several reasons:

- **Facilitate a rigorous examination of claims about AI risks**, to ensure that identified risks are well-supported by evidence and reasoning.
- **Increase the number of participants who are knowledgeable about potential AI risks** so that more participants (and their related institutions) will be able to respond to risks in a well-informed way.
- **Increase common knowledge about the risks**, i.e., ensure that participants know that other participants share their concerns. Doing so could reduce dangerous misperceptions, such as if some participants believe that other participants would behave in a reckless way because these other participants are not aware of particular risks.

Topics will have a low risk of proliferation if they do not involve information about safety risks that is not already in the public domain. Such discussions might overall *reduce* proliferation. For example, if discussing risks from AI causes participants to be more concerned about how AI could be misused, then these participants might lobby harder for measures to prevent these systems from proliferating to actors that would misuse them. Similarly, topics will have a low risk of unintended capabilities disclosure if they do not involve non-public information about different actors' AI capabilities.

Lower risk topics about technical safety risks

Pathways to catastrophic AI accidents or misuse. Discussing viewpoints on why advanced AI systems might be extremely dangerous could be helpful for increasing

[49] See the introduction of this issue brief for various examples of influential figures in the U.S. and China expressing concern about AI risks.

consensus about the true level of risk. Bengio et al. provide a good example of what kinds of points could be discussed.[50] It might also be helpful for participants to discuss possible risks in a more detailed way, such as with detailed discussions of specific ways in which AI systems could become misaligned (i.e., pursue unintended goals), e.g., via "goal misgeneralization."[51]

Implications of recent technical research about safety. By discussing recent technical safety research, participants could come to a better understanding of what this research implies for efforts to reduce AI risks and have this understanding be common knowledge. For example, researchers have demonstrated that deceptive language models would remain deceptive even after existing safety techniques are applied to them.[52] This finding has important implications for what responses would be needed to address risks that involve deceptive AI systems, such as that developers should not rely exclusively on existing safety techniques.

The discussions highlighted above could make use of just published research. In this case, the marginal effect on proliferation or unintended capabilities disclosure would presumably be low. However, these discussions might be more productive if participants can also reference non-public information, such as describing novel safety concerns that they have observed in their non-public AI systems. Such discussions would presumably have a higher likelihood of contributing to proliferation or unintended capabilities disclosure. For example, it might be difficult to discuss safety issues that arose in a non-public system without revealing some non-public information about that system's technical characteristics and capabilities.

Discussions of which pathways to AI catastrophes are most concerning might also contribute to unintended capabilities disclosure. For example, if some participants mainly describe pathways that involve AI-enabled cyberattacks, other participants might—correctly or incorrectly—conclude that the original participants have seen evidence of improved AI cyber capabilities.

Technical safety methods

A key part of reducing risks from advanced AI systems is developing technical approaches for making the systems safer. Track IIs might be helpful for sharing insights about how to implement technical measures to reduce AI risks.

[50] Bengio et al., "Managing AI Risks in an Era of Rapid Progress."
[51] For a detailed discussion of goal misgeneralization, see Shah et al., "Goal Misgeneralization."
[52] Hubinger et al., "Sleeper Agents."

A possible objection to discussing technical safety with geopolitical rivals is that safety techniques can often be used to also make AI systems more useful.[53] This phenomenon is sometimes referred to as safety techniques having "capabilities externalities."[54] Transferring insights about safety techniques with high capabilities externalities is likely to contribute to proliferation since this information could help others to build more advanced offensive or dual-use AI systems.

However, there are some directions in technical safety research that are unlikely to contribute to increased usefulness of AI systems; sharing technical information from these areas is unlikely to contribute significantly to proliferation. We highlight several of these directions in the box.[55] Ultimately, however, many insights about technical safety are likely to also contribute at least somewhat to proliferation. We would be keen to see further analysis of how to weigh up the (potentially significant) advantages and disadvantages of sharing such information.

Lower-risk topics relating to technical safety

Multi-agent safety:[56] This area focuses on ensuring AI systems remain safe when interacting with other AI systems. Failures in multi-agent interactions, even between relatively simple algorithms, have already caused real-world harm, as in the case of the 2010 "flash crash" in financial markets. Collaboration between the U.S. and China on multi-agent safety techniques could be an especially valuable type of technical collaboration. It may be difficult for actors to adequately address these issues individually because multi-agent safety failures might result from the interactions of AI systems controlled by many different actors.

- Research into multi-agent safety might be unlikely to significantly advance AI capabilities because the focus is on improving the interactions between systems rather than the capabilities of individual systems. That said, sharing this information might contribute to unintended capabilities disclosure, inasmuch as the techniques depend on the specific details about the AI systems where they are applied.

[53] Reinforcement learning from human feedback may be an example of this. Guest, Aird, and Ó hÉigeartaigh, "Safeguarding the Safeguards," 11.
[54] Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 8.
[55] Delaney and Guest, forthcoming; Guest, Aird, and Ó hÉigeartaigh, "Safeguarding the Safeguards," 13; Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 7–9, 15–20.
[56] Critch and Krueger, "AI Research Considerations for Human Existential Safety (ARCHES)," 21–23. The framing of "cooperative AI" focuses on many of the same questions as "multi-agent safety." Dafoe et al., "Open Problems in Cooperative AI"; Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 19.

Power aversion:[57] Techniques in this area would aim to address a potential hazard from advanced AI systems; these systems might attempt to obtain a lot of power over the world, such as by acquiring large amounts of money, even when not directed to do so. This is because obtaining power can be helpful for a variety of longer-term goals.[58]

- Sharing technical insights about power aversion is unlikely to proliferate capabilities; power-averse AI systems might be less capable, e.g., in that they would be constrained when carrying out tasks that require a lot of influence.[59] However, as before, sharing this information might contribute to unintended capabilities disclosure if the techniques depend on specific details about the relevant AI systems.

Anomaly detection:[60] This area focuses on identifying when AI systems are behaving in anomalous ways, such as because they have received an input that is very different from the training data. An anomaly could indicate that a new hazard has emerged, including around accident or misuse risks.

- Anomaly detection primarily serves as a safeguard rather than an enhancement to performance, so is unlikely to contribute to proliferation.[61] Once again, sharing this information might contribute to unintended capabilities disclosure, inasmuch as the techniques depend on the specific details about the AI systems where they are applied.

The research areas in the box are all relevant for reducing the risks from AI accidents. There are additional techniques that might reduce the risks from AI misuse (without increasing usefulness), such as filtering out inputs if they are likely to elicit harmful content.[62] That said, actors in the U.S. and China will sometimes have dramatically different views of what

[57] Carlsmith, "Is Power-Seeking AI an Existential Risk?," 15–31; Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 17.
[58] There are formal proofs that various kinds of AI systems would attempt to seek power and examples of this tendency in toy environments. However, empirical examples of AI power-seeking currently are not particularly compelling, perhaps because AI systems are not yet capable enough to engage in power-seeking. Hadshar, "A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking," 10–12.
[59] Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 5.
[60] Hendrycks et al., "Unsolved Problems in ML Safety," 5; Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 4, 15–16.
[61] Hendrycks and Mazeika, "X-Risk Analysis for AI Research," 28–30. Additionally, in some cases, anomaly detection could even introduce friction by generating false positives or causing delays as the system evaluates anomalies.
[62] Clifford, "Preventing AI Misuse."

constitutes misuse. For example, the Chinese government requires generative AI systems to uphold core socialist values;[63] few people in the U.S. would consider it misuse to elicit content that does not meet this requirement. Before sharing misuse-mitigation techniques, it would be important to consider what kinds of "misuse" they could and would be used to mitigate and whether this would be beneficial overall.

There are also some areas of technical safety research that make AI systems both safer and more capable. Examples include details of techniques to enhance human feedback given during AI training.[64] Sharing these insights might reduce AI risks but at the cost of some AI proliferation.

Model safety evaluations

Model safety evaluations for extreme risks aim to identify whether an AI system has dangerous capabilities, as well as its propensity to use these capabilities for harm.[65] In track II discussions, participants could discuss governance-related and technical details of how to use evaluations to reduce risks from advanced AI systems. Disseminating information about how to perform high-quality safety evaluations would presumably reduce accident risk because few actors would choose to develop or deploy an AI system that they know is dangerous.

Model safety evaluations test for two kinds of dangerous capabilities:[66]

- **Offensive capabilities**, such as knowledge about how to build bioweapons.
- **Evasive capabilities**, i.e., capabilities that would help a misaligned AI system to evade human oversight.[67] Examples include the ability of an AI system to copy itself onto a server not controlled by the developer.

National security concerns make it more favorable to share with rivals information about performing evasive capability evaluations than about performing offensive capability evaluations.[68] This is because capabilities evaluations often involve deliberately making an AI system more likely to exhibit the capability in question, such as with prompting or fine-tuning.[69] Countries might, understandably, be reluctant to share information with rivals about how to

[63] "Interim Measures for the Management of Generative Artificial Intelligence Services."
[64] Delaney and Guest, forthcoming.
[65] Shevlane et al., "Model Evaluation for Extreme Risks."
[66] Shevlane et al., 4–5.
[67] We use the term "evasive capabilities" for concision. The original paper does not give this category a clear name.
[68] If a rival seemed sufficiently likely to cause an AI accident, and sufficiently unlikely to use offensive AI capabilities, then it might be beneficial for national security to share information about offensive capability evaluations. We leave out of scope an analysis of how likely this situation is.
[69] Shevlane et al., "Model Evaluation for Extreme Risks," 13.

acquire offensive capabilities.[70] In contrast, it is harder to imagine scenarios where a rival could directly weaponize the ability of an AI system to evade any human oversight against the original country. As a result, we focus here specifically on evasive capabilities evaluations.

Lower risk topics for model safety evaluations

Governance frameworks for model safety evaluations. Participants could discuss principles and practices for conducting model safety evaluations in a way that surfaces potential risks as thoroughly as possible. Key topics could include establishing standards for auditors to be given a sufficient level of access to the AI system, designing incentive structures that encourage highly rigorous and independent audits, and fostering a culture of proactively hunting for potential concerns, even if they seem unlikely.[71]

Technical details about carrying out evasive capabilities evaluations. For example, Kinniment et al. describe several best practices for carrying out evaluations, such as taking into account potential advances in fine-tuning and scaffolding when assessing a model's capabilities.[72] Track II discussions could be valuable for sharing these kinds of technical insights, either from the literature or that participants have found but not published.

Although sharing best practices for evasive capabilities evaluations is lower risk than sharing best practices for offensive capabilities evaluations, it might still pose some risks relating to proliferation and unintended capabilities disclosure.

Sharing information about evasive capabilities evaluations may pose some proliferation risks. Countries might be reasonably concerned about their rivals gaining insights that make it easier to build AI systems that can evade human oversight, even if rivals' incentives to develop such systems are unclear. Moreover, safety-motivated techniques for increasing evasive capabilities could also be useful for enhancing AI capabilities more broadly, including offensive and

[70] Additionally, countries might also want a situation where their rivals struggle to measure these rivals' level of offensive capabilities. This ignorance might make it harder for rivals to plan to use their offensive capabilities.
[71] Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits"; Hobbhahn and Scheurer, "We Need a Science of Evals."
[72] As another example, the researchers describe simulating the results of some AI actions that would impact the real world, such as sending phishing emails. Using simulations makes it easier to study the behavior of AI systems without causing harm. Kinniment et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks."

dual-use capabilities. For instance, METR's work on evaluating AI systems' ability to autonomously replicate may advance the frontier of assessing an AI's capacity to act independently in the world.[73] These kinds of evaluation methods might be helpful for anyone attempting to design better AI agents, regardless of what they would be used for or how focused the developer is on safety.

Unintended capabilities disclosure is another potential risk when sharing information about evaluations. Some evaluation approaches may be specific to certain AI architectures, so discussing those approaches could reveal information about what architectures are being used. Additionally, some findings on best practices for conducting evaluations may only be relevant for AI systems that have reached a particular level of capability, including dual-use capabilities. Sharing such findings would imply that one possesses an AI system at that capability level.

[73] Kinniment et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks."

Appendix

Capabilities overestimates

We claimed above that actors might overestimate the level of offensive or dual-use AI capabilities of their rivals.[74] These "capabilities overestimates" seem particularly concerning from a perspective of race dynamics. If actors race against each other to develop advanced AI systems, overestimates might make them race, not just against rivals' actual level of capabilities, but against rivals' (higher) *apparent* level of capabilities.

Unintended capabilities disclosure will increase the amount of information that is available about the extent of different actors' capabilities. It is unclear whether this additional information would decrease the likelihood of capabilities overestimates. In the absence of sufficient information, estimates about rivals' AI capabilities could plausibly be either too high or too low. If the estimate is too high, then additional information will presumably (somewhat) correct an overestimate. If the estimate is too low, then additional information will presumably (somewhat) correct an underestimate.[75]

The disclosure of information about AI capabilities levels can happen via many mechanisms, not just via leaks at track IIs. For example, AI developers often publish their research or present it at conferences, and there are benchmarks to compare the capabilities of different AI systems. It is unclear whether leaks at track IIs are particularly likely to cause overestimates. On the one hand, it might be particularly difficult to verify claims during live discussions, and track II participants might allude to secret projects with capabilities that are supposedly particularly impressive.[76] On the other hand, the potential for reducing misconceptions is sometimes described as a key benefit of track IIs.[77]

---

[74] Thank you to Christian Ruhl and Saad Siddiqui for input that contributed substantially to the ideas in this appendix.

[75] In the case of states, we speculate that (in the absence of sufficient information) *over*estimates are more likely. This is because national security apparatuses often make "worst case assumptions" about potential security threats from rivals. Tang, "Fear in International Politics: Two Positions," 453–54. That said, there are some reasons to expect the opposite. For example, many governments seem to have been surprised by how impressive ChatGPT was and AlphaGo is sometimes described as a "Sputnik moment" for China, implying that Chinese actors had been underestimating the potential of AI. Department for Science, Innovation and Technology, "Introducing the AI Safety Institute"; Lee, "China's Sputnik Moment and the Sino American Battle for AI Supremacy."

[76] For a more detailed discussion of exaggerated AI capabilities in the context of international competition, see Cummings, "The AI That Wasn't There: Global Order and the (Mis)Perception of Powerful AI.".

[77] Kerrigan, Grek, and Mazarr, *The United States and China? Designing a Shared Future: The Potential for Track 2 Initiatives to Design an Agenda for Coexistence*, 11–12.

We would be interested in further research on the question of whether track IIs are likely to (significantly) contribute to capabilities overestimates and, if so, what can be done to mitigate this risk. For example, perhaps debriefs after a track II would help participants to notice if they have a skewed impression of the other side's capabilities.

Chinese interest in non-proliferation measures

As highlighted in the body, the position of the Chinese government on AI proliferation is not yet clear.

- On the one hand, the Chinese government has expressed significant concern about the risks of AI misuse by criminals or terrorists.[78] Non-proliferation measures would be one way to address these concerns. There is also precedent of the U.S. and China cooperating to reduce the proliferation of dangerous technologies, even as they compete between themselves in those same technologies. For example, the U.S. and China are both members of the NPT and Nuclear Suppliers Group, two regimes to reduce nuclear proliferation.[79]
- On the other hand, China's "Global AI Governance Initiative" emphasizes diffusing AI development capabilities. The Ministry of Foreign Affairs' statement about the initiative states: "We should uphold the principles of mutual respect, equality, and mutual benefit in AI development. All countries, regardless of their size, strength, or social system, should have equal rights to develop and use AI."[80]
- China has also frequently criticized the U.S.-led controls on exports of AI-relevant hardware to many countries, a non-proliferation measure.[81] That said, this criticism is hardly surprising given that China is the primary target of these controls.[82]

[78] Concordia AI, "State of AI Safety in China," 31–32.
[79] Note that the U.S. argues that China has not always complied with these obligations, e.g., with dubious nuclear exports to Pakistan. Kerr, "Chinese Nuclear and Missile Proliferation," 2.
[80] Wu Zhaohui, China's vice minister of science and technology, made similar remarks at the UK AI Safety Summit. "Global AI Governance Initiative"; "China, U.S. and EU Sign Milestone Declaration to Teamwork in AI Safety."
[81] "China Lashes out at Latest U.S. Export Controls on Chips"; "Foreign Ministry Press Conference on January 8, 2024."
[82] Dohmen and Feldgoise, "A Bigger Yard, A Higher Fence."

Acknowledgements

We are grateful to the following people for discussion and input: Mauricio Baker, Karson Elmgren, Fynn Heide, Christian Ruhl, Saad Siddiqui, and the participants of an internal IAPS seminar. Mistakes and opinions are our own. We are also grateful to Maya Deutchman for copyediting.

Bibliography

AI Safety Fundamentals. "Avoiding Extreme Global Vulnerability as a Core AI Governance Problem," November 8, 2022. https://aisafetyfundamentals.com/blog/global-vulnerability/.

Allen, Gregory C. "Choking off China's Access to the Future of AI." CSIS, November 10, 2022. https://www.csis.org/analysis/choking-chinas-access-future-ai.

Anderljung, Markus, and Julian Hazell. "Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted?" arXiv, March 29, 2023. http://arxiv.org/abs/2303.09377.

Anthropic. "Anthropic's Responsible Scaling Policy, Version 1.0," September 19, 2023. https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

AP News. "China Lashes out at Latest U.S. Export Controls on Chips," October 8, 2022. https://apnews.com/article/technology-business-china-global-trade-47eed4a9fa1c2f51027ed12cf929ff55.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI & SOCIETY* 31, no. 2 (May 2016): 201–6. https://doi.org/10.1007/s00146-015-0590-y.

Baker, Mauricio. "Nuclear Arms Control Verification and Lessons for AI Treaties." arXiv, April 8, 2023. http://arxiv.org/abs/2304.04123.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, et al. "Managing AI Risks in an Era of Rapid Progress." arXiv, October 26, 2023. https://doi.org/10.48550/arXiv.2310.17688.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." arXiv, February 20, 2018. https://doi.org/10.48550/arXiv.1802.07228.

Carlsmith, Joseph. "Is Power-Seeking AI an Existential Risk?" arXiv, June 16, 2022. http://arxiv.org/abs/2206.13353.

Casper, Stephen, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, et al. "Black-Box Access Is Insufficient for Rigorous AI Audits." arXiv, January 25, 2024. https://doi.org/10.48550/arXiv.2401.14446.

Center for AI Safety. "Statement on AI Risk," May 30, 2023. https://www.safe.ai/statement-on-ai-risk.

Center for Human-Compatible Artificial Intelligence (CHAI). "Prominent AI Scientists from China and the West Propose Joint Strategy to Mitigate Risks from AI," October 31, 2023. https://humancompatible.ai/news/2023/10/31/prominent-ai-scientists-from-china-and-the-west-propose-joint-strategy-to-mitigate-risks-from-ai/.

Center For International Security And Strategy, Tsinghua University. "The China-U.S. Track II Dialogue on Artificial Intelligence and International Security Interim Report," April 6, 2024. https://ciss.tsinghua.edu.cn/info/banner/7041.

CGTN. "China, U.S. and EU Sign Milestone Declaration to Teamwork in AI Safety," November 2, 2023. https://news.cgtn.com/news/2023-11-02/China-U-S-and-EU-sign-milestone-declaration-to-teamwork-in-AI-safety-1ooWSuT1RoQ/index.html.

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

Christiano, Paul. "Thoughts on the Impact of RLHF Research." Alignment Forum, January 25,

2023. https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research.
Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences." arXiv, February 17, 2023. http://arxiv.org/abs/1706.03741.
Clifford, Ben. "Preventing AI Misuse: Current Techniques." Centre for the Governance of AI, December 17, 2023. https://www.governance.ai/post/preventing-ai-misuse-current-techniques.
Coe, Andrew J., and Jane Vaynman. "Why Arms Control Is So Rare." *American Political Science Review* 114, no. 2 (May 2020): 342–55. https://doi.org/10.1017/S000305541900073X.
Concordia AI. "State of AI Safety in China," October 2023. https://concordia-ai.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf.
Concordia AI. "The State of China-Western Track 1.5 and 2 Dialogues on AI." AI Safety in China, August 24, 2023. https://aisafetychina.substack.com/p/the-state-of-china-western-track.
Critch, Andrew, and David Krueger. "AI Research Considerations for Human Existential Safety (ARCHES)." arXiv, May 29, 2020. http://arxiv.org/abs/2006.04948.
Cummings, Mary (Missy). "The AI That Wasn't There: Global Order and the (Mis)Perception of Powerful AI." In *Policy Roundtable: Artificial Intelligence and International Security*. Austin, TX: Texas National Security Review, 2020. https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/#essay2.
Dafoe, Allan, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. "Open Problems in Cooperative AI." arXiv, December 15, 2020. http://arxiv.org/abs/2012.08630.
Delaney, Oscar, and Oliver Guest. "AI Risk-Reduction Efforts Overlooked by Industry," forthcoming.
Department for Science, Innovation and Technology. "Introducing the AI Safety Institute." GOV.UK, November 2023. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.
Dohmen, Hanna, and Jacob Feldgoise. "A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports." Center for Security and Emerging Technology, December 4, 2023. https://cset.georgetown.edu/article/bis-2023-update-explainer/.
Ee, Shaun, and Joe O'Brien. "Putting New AI Lab Commitments in Context." Centre for the Governance of AI, August 4, 2023. https://www.governance.ai/post/putting-new-ai-lab-commitments-in-context.
Emery-Xu, Nicholas, Andrew Park, and Robert Trager. "Uncertainty, Information, and Risk in International Technology Races." *Journal of Conflict Resolution*, November 17, 2023, 00220027231214996. https://doi.org/10.1177/00220027231214996.
Gleave, Adam. "AI Safety in a World of Vulnerable Machine Learning Systems." FAR AI, March 5, 2023. https://far.ai/post/2023-03-safety-vulnerable-world/.
Godek, Sarah. "Why China's Export Controls on Germanium and Gallium May Not Be Effective." Stimson Center, July 19, 2023. https://www.stimson.org/2023/why-chinas-export-controls-on-germanium-and-gallium-may-not-be-effective/.

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

Guest, Oliver, Michael Aird, and Fynn Heide. “International AI Safety Dialogues: Benefits, Risks, and Best Practices.” Institute for AI Policy and Strategy, December 14, 2023. https://www.iaps.ai/research/safeguarding-the-safeguards.

Guest, Oliver, Michael Aird, and Seán Ó hÉigeartaigh. “Safeguarding the Safeguards: How Best to Promote AI Alignment in the Public Interest.” Institute for AI Policy and Strategy, December 15, 2023. https://www.iaps.ai/research/safeguarding-the-safeguards.

Gupta, Abhishek, Camylle Lanteigne, and Victoria Heath. “Report Prepared by the Montreal AI Ethics Institute (MAIEI) on Publication Norms for Responsible AI.” arXiv, October 4, 2020. https://doi.org/10.48550/arXiv.2009.07262.

Hadshar, Rose. “A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking.” arXiv, October 27, 2023. http://arxiv.org/abs/2310.18244.

Hass, Ryan, and Colin Kahl. “Laying the Groundwork for US-China AI Dialogue.” Brookings, April 5, 2024. https://www.brookings.edu/articles/laying-the-groundwork-for-us-china-ai-dialogue/.

Heide, Fynn. “Beijing Policy Interest in General Artificial Intelligence Is Growing,” June 8, 2023. https://www.governance.ai/post/beijing-policy-interest-in-general-artificial-intelligence-is-growing.

Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. “Unsolved Problems in ML Safety.” arXiv, June 16, 2022. http://arxiv.org/abs/2109.13916.

Hendrycks, Dan, and Mantas Mazeika. “X-Risk Analysis for AI Research.” arXiv, September 20, 2022. https://doi.org/10.48550/arXiv.2206.05862.

Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. “International Institutions for Advanced AI.” arXiv, July 11, 2023. http://arxiv.org/abs/2307.04699.

Hobbhahn, Marius, and Jeremy Scheurer. “We Need a Science of Evals.” Apollo Research, 2024. https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.

Horowitz, Michael, and Paul Scharre. “AI and International Stability: Risks and Confidence-Building Measures.” Washington, DC: Center for a New American Security, January 12, 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. “Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training.” arXiv, January 17, 2024. http://arxiv.org/abs/2401.05566.

Imbrie, Andrew, and Elsa Kania. “AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement.” Center for Security and Emerging Technology, December 2019. https://doi.org/10.51593/20190051.

Jervis, Robert. “Arms Control, Stability, and Causes of War.” *Political Science Quarterly* 108, no. 2 (June 1, 1993): 239–53. https://doi.org/10.2307/2152010.

Jones, Peter L. *Track Two Diplomacy in Theory and Practice*. Stanford, California: Stanford University Press, 2015.

Kerr, Paul K. “Chinese Nuclear and Missile Proliferation.” IN FOCUS. Washington, DC: Congressional Research Service, October 24, 2023.

https://crsreports.congress.gov/product/pdf/IF/IF11737.
Kerrigan, Amanda, Lydia Grek, and Michael J. Mazarr. *The United States and China? Designing a Shared Future: The Potential for Track 2 Initiatives to Design an Agenda for Coexistence*. Santa Monica, CA: RAND Corporation, 2023. https://doi.org/10.7249/RRA2850-1.
Kinniment, Megan, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, et al. "Evaluating Language-Model Agents on Realistic Autonomous Tasks." arXiv, January 4, 2024. https://doi.org/10.48550/arXiv.2312.11671.
Kissinger, Henry A., and Graham Allison. "The Path to AI Arms Control." *Foreign Affairs*, October 13, 2023. https://www.foreignaffairs.com/united-states/henry-kissinger-path-artificial-intelligence-arms-control.
Lamberth, Megan, and Paul Scharre. "Arms Control for Artificial Intelligence." *Texas National Security Review* 6, no. 2 (2023): 95–110. https://doi.org/10.26153/TSW/46142.
Lee, Kai-Fu. "China's Sputnik Moment and the Sino American Battle for AI Supremacy." Asia Society, December 3, 2019. https://asiasociety.org/magazine/article/chinas-sputnik-moment-and-sino-american-battle-ai-supremacy.
Leike, Jan. "Distinguishing Three Alignment Taxes." Musings on the Alignment Problem, December 19, 2022. https://aligned.substack.com/p/three-alignment-taxes.
Maas, Matthijs M., and José Jaime Villalobos Ruiz. "International AI Institutions: A Literature Review of Models, Examples, and Proposals." *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4579773.
MacCarthy, Mark. "The US and Its Allies Should Engage with China on AI Law and Policy." Brookings, October 19, 2023. https://www.brookings.edu/articles/the-us-and-its-allies-should-engage-with-china-on-ai-law-and-policy/.
Ministry of Foreign Affairs of the People's Republic of China. "Foreign Ministry Spokesperson Mao Ning's Regular Press Conference on January 8, 2024," January 8, 2024. https://www.fmprc.gov.cn/eng/xwfw_665399/s2510_665401/2511_665403/202401/t20240108_11219851.html.
Ministry of Foreign Affairs of the People's Republic of China. "Global AI Governance Initiative," October 20, 2023. https://www.mfa.gov.cn/eng/wjdt_665385/2649_665393/202310/t20231020_11164834.html.
OpenAI. "Preparedness," December 18, 2023. https://openai.com/safety/preparedness.
Poli, Michael, Stefano Massaroli, Eric Nguyen, Dan Fu, Tri Dao, Stephen A. Baccus, Yoshua Bengio, Stefano Ermon, and Chris Ré. "Hyena Hierarchy: Towards Larger Convolutional Language Models," March 7, 2023. https://hazyresearch.stanford.edu/blog/2023-03-07-hyena.
Räuker, Tilman, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks." arXiv, August 18, 2023. https://doi.org/10.48550/arXiv.2207.13243.
Safe AI Forum. "International Dialogues on AI Safety." Accessed April 9, 2024. https://idais.ai/.
Schneider, Jordan, and Ryan Hauser. "Pottinger on Trump 2.0." ChinaTalk, March 20, 2023. https://www.chinatalk.media/p/pottinger-on-trump-20.
Schuett, Jonas, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. "Towards Best Practices in AGI Safety and Governance: A

Survey of Expert Opinion." arXiv, May 11, 2023. http://arxiv.org/abs/2305.07153.
Select Committee on the CCP. "Select Committee Republicans Issue Demands For Xi Jinping Ahead of Meeting with President Biden," November 9, 2023. http://selectcommitteeontheccp.house.gov/media/press-releases/select-committee-republicans-issue-demands-xi-jinping-ahead-meeting-president.
Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." arXiv, November 2, 2022. https://doi.org/10.48550/arXiv.2210.01790.
Sheehan, Matt. "What the U.S. Can Learn From China About Regulating AI." Foreign Policy, February 28, 2024. https://foreignpolicy.com/2023/09/12/ai-artificial-intelligence-regulation-law-china-us-schumer-congress/.
Shevlane, Toby. "Structured Access: An Emerging Paradigm for Safe AI Deployment." arXiv, April 11, 2022. http://arxiv.org/abs/2201.05159.
Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, et al. "Model Evaluation for Extreme Risks." arXiv, May 24, 2023. http://arxiv.org/abs/2305.15324.
Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, et al. "Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings." arXiv, August 3, 2023. https://doi.org/10.48550/arXiv.2308.00862.
Tang, Shiping. "Fear in International Politics: Two Positions." *International Studies Review* 10, no. 3 (2008): 451–71.
The White House. "President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence," October 30, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.
Trager, Robert F., Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, et al. "International Governance of Civilian AI: A Jurisdictional Certification Approach." Oxford: Oxford Martin AI Governance Initiative in partnership with the Centre for the Governance of AI, August 2023. governance.ai/research-paper/international-governance-of-civilian-ai.
Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. "Dual Use of Artificial-Intelligence-Powered Drug Discovery." *Nature Machine Intelligence* 4, no. 3 (March 7, 2022): 189–91. https://doi.org/10.1038/s42256-022-00465-9.
Wang, Yifan, and Rongsheng Zhu. "What Topics Can Be Discussed in the China-U.S. Artificial Intelligence Dialogue [中美人工智能对话可以谈些什么]." *CSIS Interpret: China, Original Work Published in World Affairs [世界知识]*, January 2024. https://interpret.csis.org/translations/what-topics-can-be-discussed-in-the-china-u-s-artificial-intelligence-dialogue/.
Webster, Graham, and Ryan Hass. "A Roadmap for a US-China AI Dialogue." Brookings, January 10, 2024. https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/.
Whittlestone, Jess, and Aviv Ovadya. "The Tension between Openness and Prudence in AI Research." arXiv, January 13, 2020. https://doi.org/10.48550/arXiv.1910.01170.
Ying, Fu, and John Allen. "Together, The U.S. And China Can Reduce The Risks From AI." NOEMA, December 17, 2020.

https://www.noemamag.com/together-the-u-s-and-china-can-reduce-the-risks-from-ai.