International AI safety dialogues

Benefits, risks, and best practices

Institute for AI Policy and Strategy (IAPS)

October 31, 2023

AUTHORS

Oliver Guest — Research Analyst, IAPS

Michael Aird — Acting Co-director, IAPS

Fynn Heide — Research Scholar, Centre for the Governance of AI

Abstract

Events that bring together stakeholders from a range of countries to talk about AI safety (henceforth "safety dialogues") are a promising way to reduce large-scale risks from advanced AI systems. The goal of this report is to help safety dialogue organizers make these events as effective as possible at reducing such risks. We first identify “best practices” for organizers, drawing on research about comparable past events, literature about track II diplomacy, and our experience with international relations topics in AI governance. We then identify harmful outcomes that might result from safety dialogues, and ideas for how organizers can avoid them. Finally, we overview AI safety interventions that have already been identified and that might be particularly fruitful to discuss during a safety dialogue.

Executive summary

Introduction

(To section)

International AI safety dialogues ("safety dialogues") are events that bring together stakeholders from a range of countries to talk about AI safety. With this report, we aim to help organizers of safety dialogues to make these events as effective as possible at reducing large-scale risks from advanced AI.

We predominantly focus on a specific type of safety dialogue with the following properties:

- Focus on catastrophic or even existential risks from highly advanced yet "misaligned" AI systems.
- Include participants from (at least) US and Chinese institutions.
- Include only participants that are not official representatives of their governments, along the lines of "track II" diplomacy.
- Mainly include participants with a technical AI background.

That said, we expect that many of our findings would apply to safety dialogues more broadly, and that it could be valuable for a range of different and complementary safety dialogues to occur.[1]

In this report, we first highlight best practices for safety dialogue organizers. We then discuss downsides that organizers should attempt to avoid. Finally, we describe proposed AI safety interventions that might be particularly fruitful to discuss at safety dialogues.

Best practices for organizers

(To section)

Drawing on research about comparable past events, literature about track II diplomacy, and our experience with international relations topics in AI

[1] Note, however, that some of our recommendations might be inappropriate for some other types of safety dialogue. As an example, intergovernmental safety dialogues, such as the upcoming UK AI Safety Summit, might inevitably involve more negotiation than we suggest would be ideal for our specific type of safety dialogue.

governance, we identify the following best practices that organizers can follow to increase the likelihood that safety dialogues successfully contribute to reducing large-scale AI risks.

Culture of the safety dialogues

- **Make the dialogue non-partisan** in order to avoid alienating some potential participants or outside stakeholders. In particular, safety dialogues should not be biased for or against the US or China, and they should be seen as such. Organizers can contribute to this by seeking non-partisan sources to fund the safety dialogue. The "truth-seeking" spirit described immediately below may also be helpful for reducing this kind of partisanship.
- **Promote a spirit of collaborative truth-seeking among participants**, rather than more adversarial framings such as negotiations. For example, organizers could encourage participants to work together to better understand relevant technical questions. Similarly, organizers could encourage participants to work together in areas where they can agree, such as technical questions relating to alignment, even if agreement on other topics, such as US-China relations, is harder.
- **Create high-trust relationships between the participants**. For example, organizers can encourage social interactions between participants. Additionally, organizers should try to ensure that participants understand the limits of each other's influence, so that trust is not damaged if participants notice another participant's organization doing something that is inconsistent with the views expressed by that participant.
- **Create high-trust relationships between the participants and facilitators**. For example, facilitators can demonstrate that they care about the perspectives of individual participants, such as by avoiding seeming judgemental towards particular perspectives.

Communicating about safety dialogues to outsiders

- **Maintain confidentiality about what was said by whom**. A "Chatham House" rule might make it easier for participants to speak freely, contributing to high-quality discussion.
- **Consider maintaining confidentiality about who is attending**. There are advantages to this approach, such as making it easier for some participants to attend and speak freely. That said, there are also disadvantages, such as potentially reducing the perceived credibility of the safety dialogues. Organizers should weigh the advantages and disadvantages for their specific case. One "best of both worlds" approach may be to publicly announce only some of the participants' names.

- **Consider publishing a readout after the dialogue**. Depending on participants' views, a readout could create common knowledge about participants' concerns and the fact that experts in both the US and China are concerned about AI risks. This might motivate further AI safety work and help avoid a scenario where US and Chinese actors both avoid implementing AI safety measures because of a belief that these will not be reciprocated.

Content of the event

- **Facilitators should provide inputs to encourage participants down a productive path**, such as noting ways in which participants may be talking past each other, and providing relevant empirical information (e.g., technical findings about AI alignment) that might not be known to all participants.
- **Sometimes split participants into working groups**, particularly when the main group is deadlocked. This might make it easier to find a position that all participants can share.

Selecting participants to invite

- **Choose participants who will engage constructively**, e.g., because they are keen to work with others to improve AI safety and will follow the norms of the safety dialogue.
- **Consider including participants from a range of countries**. While this has some potential downsides, we expect that this would generally be helpful, even if the main goal of the summit is to improve AI safety efforts specifically in the US and China. For example, including third countries might make participants less likely to view the event through the frame of zero-sum competition between the US and China. Avoiding this frame might make it easier for participants to find areas where they or their institutions can cooperate to promote safety.
- **Consider the right level of participant "turnover" between dialogues**, if the dialogues are recurring. On the one hand, a high turnover rate would increase the number of participants that meet each other, potentially creating a higher number of valuable relationships. On the other hand, repeated interactions between the same participants could be helpful for creating particularly strong relationships. We expect that the right balance between these considerations will vary from case to case.

Logistical details

- **Choose a suitable location, such as by considering accessibility to participants, comfort, and a relatively "neutral" location**. For example, Singapore and Switzerland each have fairly good relations with both the US and China.

- **Reduce language barriers**, such as by providing translators or by creating glossaries of key AI safety terms. The process of convening experts to create these glossaries may itself be a helpful basis for cooperation on AI safety.

Harmful outcomes to avoid

(To section)

Safety dialogues could lead to harmful outcomes, reducing the overall value of the safety dialogue, or even causing the safety dialogue to do more harm than good. We overview possible harmful outcomes that are particularly concerning and suggest ways for organizers to reduce these risks.

Promoting interest in AI capabilities disproportionately, relative to AI safety

- Discussions about the risks from powerful AI systems might backfire if participants focus too strongly on the potential benefits of powerful AI and too little on the risks. Similarly, safety dialogues might backfire if participants become focused on competition around developing powerful new systems, rather than cooperation to avoid catastrophe. These worldviews might cause participants to promote reckless AI development without sufficient attention to safety.
    - → Facilitators should aim to keep participants focused on the topic of the safety dialogue, i.e., safety concerns around AI. Additionally, it may be helpful to avoid focusing discussions on potential military applications of AI. National security is often framed in a zero-sum way, and thus is particularly likely to promote thinking about competitive strategic dynamics, as opposed to accident risks that might affect everyone.

Reducing the influence of safety concerns

- Poorly managed safety dialogues could reduce the influence of safety-focused actors, such as the people who choose to attend these dialogues. For example, US and Chinese participants meeting each other could be seen by some outsiders as improperly engaging with adversaries, hurting the careers or credibility of these participants. As another example, government officials may interpret participants' actions as unauthorized efforts to influence foreign policy, provoking backlash.

- → To prevent this, organizers can emphasize the value of cooperating (even with rivals) to reduce risks, and highlight that there have been many examples of this in the past. Additionally, organizers should consider inviting participants from a range of countries, even if the organizers are primarily aiming to promote dialogue between participants from US and Chinese institutions. Having third countries might somewhat reduce the likelihood of safety dialogues being seen through the lens of US-China geopolitical competition, potentially making them less controversial to outsiders.

Diffusing AI capabilities insights

- AI safety dialogues might inadvertently spread technical insights about how to build more powerful AI systems. This "capabilities diffusion" might be harmful in three ways. First, institutions may be less willing to allow people associated with them to participate in safety dialogues if they are worried that these people will diffuse capabilities from their institutions to potential rivals. Second, diffusion could accelerate progress at the AI capabilities frontier, leaving less time for safety preparations before the most advanced AI systems are extremely powerful. Third, diffusion could spread technical know-how among more actors, making coordination on safety more difficult and increasing the likelihood that one actor uses advanced AI in a reckless or malicious way.[2]
  - → To reduce the likelihood of capabilities diffusion, organizers could establish guidelines against sharing technical specifics, select participants who are likely to follow these guidelines, and focus discussions on conceptual issues, where capabilities diffusion is less likely.

Interventions to discuss at safety dialogues

(To section)

Discussing specific AI safety interventions at safety dialogues might be a helpful way to ground the discussion. It might also increase clarity about which interventions would be desirable, as well as technically and politically feasible. If a given intervention is desirable and feasible, safety dialogues could then contribute to implementing this intervention, such as by providing a space where details can be worked out and by helping relevant people to coordinate

[2] That said, there are also legitimate reasons to want broader access to advanced AI, such as concerns about concentration of power.

on implementing the intervention. Additionally, for any interventions that would require or benefit from international cooperation or coordination, safety dialogues might be a helpful "stepping stone" towards this. For example, safety dialogues might improve trust between people of different countries and increase international consensus about the value of a given intervention.

We describe several interventions that might be fruitful to discuss in a safety dialogue, drawing on several prominent proposals from the AI governance field.[3] We do not mean to imply, however, that safety dialogue organizers or participants should assume that these interventions are desirable or feasible; reasonable people can disagree about this. Additionally, safety dialogues will generally work best if participants have a sense of "co-creating." As such, organizers should be sure to allow individual participants to meaningfully shape the results of the discussion, even if organizers present specific potential interventions to discuss.

We describe an "overarching plan" for reducing large-scale risks from AI misalignment, consisting of licensing of cutting-edge training runs, pre-deployment safety evaluations, and tracking compute clusters. There are also various standalone measures that might reduce AI safety risks if implemented by individual AI labs or other actors. For example, labs could conduct risk assessments before deploying particularly capable models, and policymakers could improve monitoring of AI safety incidents.

[3] In particular, the "overarching plan" is similar to the measures proposed in two papers that were co-authored by many of the leading figures in those fields (Anderljung et al., 2023; Shevlane et al., 2023). Additionally, the best practices that we highlight scored highly in two recent surveys of those fields (Räuker & Aird, 2023; Schuett et al., 2023).

Table of Contents

1. Introduction

Recent events have caused widespread concern about catastrophic and even existential risks associated with advanced AI systems.[4] This concern has created significant demand for interventions that might address these risks.

International AI safety dialogues ("safety dialogues" for short) seem to us to be a promising AI safety intervention.[5] Safety dialogues is our term for events that bring together stakeholders from a range of countries to talk about AI safety. Two possible effects of safety dialogues are particularly promising. First, safety dialogues might increase concern and understanding about AI safety among AI-relevant actors around the world. This would help them to make better decisions in relation to AI safety. Second, safety dialogues may be a helpful stepping stone towards deeper international cooperation on AI safety. Given that many proposed AI safety interventions, such as treaties or a global AI regulator (Altman et al., 2023; Ho et al., 2023; Marcus & Reuel, 2023), would require significant international cooperation, such a stepping stone could be extremely valuable.

The goal of this report is to help safety dialogues to be as effective as possible for reducing large scale AI risks, including catastrophic and even existential risks. Our recommendations are primarily aimed at people who are organizing, funding, or facilitating safety dialogues ("organizers"), though some of the recommendations may also apply to safety dialogue participants.

In this report, we focus on a specific type of safety dialogue. We expect that this type would be particularly helpful at getting participants to converge on AI safety actions that their respective institutions and countries should be taking, and at encouraging these actors to take these actions. Such international networks of experts ("epistemic communities") seem to have been helpful at promoting international cooperation in previous high-stakes contexts, such as nuclear arms control (Maas, 2019, pp. 12–15; ÓhÉigeartaigh et al., 2020, pp. 581–582).

This safety dialogue has the following properties:

[4] See, for example, the release of ChatGPT and GPT-4, prominent reporting on strange behaviors by the Bing/Sydney system, and public statements organized by the Future of Life Institute and the Center for AI Safety (Hogarth, 2023; Perrigo, 2023; Vallance, 2023).

[5] By "AI safety" we mean the challenge of ensuring that advanced AI systems are safe and beneficial. This includes both technical work (e.g., ensuring that a given AI system is "aligned") as well as governance work to create desirable norms, policies, and institutions around AI development and deployment (Amodei et al., 2016, pp. 20–21; cf. Dafoe, 2018, pp. 25–33).

- **Focus on catastrophic or existential risks from AI "alignment failures."**[6] Having a somewhat narrow scope might make it easier for participants to make progress.[7] We do not mean to imply, however, that other risks from AI are unimportant or should not also be addressed, including in other safety dialogues.[8]
- **Include participants from (at least) US and Chinese institutions**. We focus on these countries because they are among the leaders in building powerful, and thus potentially dangerous, AI systems.[9] Simultaneously, they have a very strained relationship, making it less likely that safety-relevant cooperation and information-sharing would emerge between them by default. As such, successful dialogues involving the US and China seem particularly valuable.
- **Only include participants that are not official representatives of their governments**, along the lines of track II diplomacy.[10] That said, even if participants are not official government representatives, we expect that participants from Chinese institutions will on average have closer government connections than their US counterparts.[11] Although we do not think that Chinese participants should be excluded for this reason, we expect that it will influence the dynamics of safety dialogues.[12] For example, it might affect how non-Chinese participants interpret the words of their Chinese counterparts, and it may limit what Chinese participants can say.
- **Mainly include participants with a technical background** in machine learning or in AI more broadly.

[6] For more on the alignment problem in high-stakes contexts, see, e.g., Critch & Krueger (2020), Ngo et al. (2023), and Russell (2020).

[7] At the extreme, if all (possible) harms associated with AI are in scope, then participants would only have the time to engage superficially with each topic that they discuss.

[8] Other risks from AI include misuse and structural risks at the catastrophic or existential level, as well as a range of harms with less extreme stakes but that might be more likely. For taxonomies of (extreme) AI risks, see Hendrycks et al. (2023), Maham & Küspert (2023), and Zwetsloot & Dafoe (2019). See Bullock et al. (2022) for wide-ranging discussions of harms from AI.

[9] See, for example, the various metrics in chapter one of Maslej et al. (2023).

[10] In track II diplomacy, participants can still have some links to their governments (e.g., via close professional contacts) but are understood to be speaking in a personal capacity (Jones, 2015, p. 25).

[11] The Chinese government likely exerts stronger political influence over universities and companies in its jurisdiction than the US government (Heilmann, 2017, pp. 207–212; Sheehan, 2023, pp. 20–22). Additionally, the Chinese government seems to be more involved in cutting-edge AI development in its country than the US government. In particular, there arguably is no US equivalent for the Beijing Academy of Artificial Intelligence (BAAI): BAAI is a quasi-government organization, is funded by the central Chinese government and the Beijing municipal government, and is one of the leading developers of frontier AI models (Ding & Xiao, 2023, pp. 4–8).

[12] There is significant precedent for international dialogues where one group of participants has closer government connections than another. For example, the Pugwash Conferences had a comparable dynamic, with Soviet participants being more connected to their government than US participants (Lüscher, 2020, p. 121).

Despite this scoping decision, we expect that other types of safety dialogues would also be helpful and could complement the type of dialogue described here. For example, we are glad that the UK government is hosting an "AI Safety Summit" (Department for Science, Innovation and Technology, 2023). Moreover, we expect that much of this report (in particular, sections two and three) would apply to other types of safety dialogue.[13]

After the introduction, this report has three main sections. Section two focuses on "best practices" for organizing and running safety dialogues. We primarily highlight lessons that have been learned elsewhere and apply them to this new context. Section three identifies possible harmful outcomes from safety dialogues that organizers should try to avoid. It gives some thoughts on how to do so. Section four lays out various AI governance interventions that might be particularly fruitful to discuss during safety dialogues.

[13] That said, there is little discussion here of points that would be relevant for (say) the UK's AI Safety Summit but not for the type of safety dialogue that is the focus of the report.

2. Best practices for organizers

We identify "best practices" for organizers that we expect would make safety dialogues particularly effective at reducing large-scale risks from AI, first describing our research method, and then giving a series of recommendations, grouped by theme.

Method for identifying recommendations

Because there is little existing work specifically about safety dialogues, our recommendations come from three complementary "strands" of knowledge, described just below. As we have highlighted, all three strands have limitations for making recommendations about safety dialogues. As such, it is possible that organizers should deviate from our recommendations in cases where our rationale for a given best practice does not seem to apply.

Strand 1: Published work about analogous past events

In our first strand, we looked for specific past events that are analogous to safety dialogues. We were particularly interested in events that shared two key characteristics with safety dialogues. The first characteristic is that the event brings together participants from different countries – and, in particular, countries that have a strained relationship with each other. The second characteristic is that the goal of the event is to reduce risks (particularly high-stakes risks) from a powerful technology.

We found three cases that are particularly relevant:[14]

- **Pugwash Conferences during the Cold War (from 1957)**: These conferences brought together scientists from both sides of the Iron Curtain to discuss nuclear arms control (Kraft & Sachse, 2020). We focus particularly on this example because it seems more relevant than the others, and because there is an especially rich literature about it.
- **Asilomar Conference (1975)**: Asilomar brought together scientists to discuss possible hazards and suggest safety guidelines for recombinant DNA research; rDNA was seen as a dangerous new technology at the time (Grace, 2015).
- **Cross-Cultural AI Ethics and Governance Workshop Series (from 2019)**: These workshops bring together participants from Eastern and Western

[14] To find these cases, we did a non-systematic literature review, particularly looking for cases that meet our criteria. See the Appendix for other events that somewhat meet the criteria and that we considered including.

countries to improve coordination on AI, but do not center on catastrophic or existential risks (Center for Long-term Artificial Intelligence, 2022; ÓhÉigeartaigh et al., 2020).

We provide additional detail on all three cases in the Appendix. This Appendix also includes discussion of disanalogies between these cases and safety dialogues, as well as additional cases that we considered including.

Strand 2: Published work about track II diplomacy

In our second strand, we looked in the literature for best practices that have been identified in relation to track II diplomacy.[15] Following the original definition, we use "track II diplomacy" to mean "unofficial, informal interaction between members of adversarial groups or nations with the goals of developing strategies, influencing public opinion, and organizing human and material resources in ways that might help resolve the conflict" (Jones, 2015, p. 9).[16] We mostly draw on Burgess & Burgess (2010) and Jones (2015). Both sources are primarily handbooks for running track II processes, written by experienced practitioners.

Track II events and safety dialogues have important similarities. For example, both bring together non-government individuals from countries with strained relationships. That said, the track II processes that Burgess & Burgess and Jones have in mind are primarily focused on resolving violent conflict, particularly between subnational groups. This is disanalogous to safety dialogues.

Strand 3: Tacit knowledge from the AI governance field

Finally, we used our own reasoning and experience of the AI Governance field to create recommendations. We specialize in AI Governance as it relates to International Relations and/or China Studies. That said, we have limited experience with safety dialogues in particular.

"Best practice" recommendations

Drawing on these three strands, we identified the following best practices for safety dialogues. The best practices can be read in any order.

[15] Future research could extend this "strand" by trying to identify best practices from track 1.5 diplomacy, i.e., meetings that involve both government officials (who participate in an unofficial capacity) and non-governmental experts (Staats et al., 2019). That said, our impression is that the literature about track 1.5 diplomacy is thinner than the literature about track II diplomacy.

[16] Note that "conflict" here has a broader meaning than armed violence and also includes, for example, different groups perceiving their interests to be incompatible.

As highlighted before, we focused on a particular type of safety dialogue when identifying these best practices; we are primarily writing about safety dialogues that focus on large-scale risks from AI misalignment, and that include participants from US and Chinese institutions who are not official representatives of their governments. Although some best practices will apply to safety dialogues in general, we expect that some best practices will not apply to safety dialogues that are, for example, primarily between governments.[17]

Culture of the safety dialogues

By "culture," we primarily mean the norms and understandings that are shared by organizers and participants of safety dialogues. We make several recommendations here for the culture that organizers should attempt to create.

Make the dialogue non-partisan

Safety dialogues should be, and be correctly seen as, non-partisan. In particular, safety dialogues that include participants from institutions in the US and China would ideally not be seen in these countries as biased for or against the US and/or China. Additionally, safety dialogues would ideally not be associated with a particular position on the political spectrum, such as Democratic or Republican in the US context.

If safety dialogues are (seen as) biased against a particular actor, such as a particular group or country, that actor is less likely to engage with

[17] Here is a brief list of some ways in which these best practices might not apply if the safety dialogue is primarily between governments:

- Advice about what types of participants to select might not apply if the safety dialogue is between governments. In an intergovernmental context, we expect that organizers would generally be able to choose which governments (or maybe which government departments) to invite, but that the invited government would generally choose which specific people to send.
- If the safety dialogue is between governments, then it might naturally look like a negotiation, despite our view that the safety dialogue that we focus on should have an atmosphere of collaborative truth-seeking.
- The "Chatham House" norm that we propose is unusual in official intergovernmental meetings.
- Various logistical details will presumably be different if the dialogue is between governments, e.g. because governments already have infrastructure in place to support summits and because the security requirements might be higher.

The best practices here might also not apply to safety dialogues that differ in other ways, such as with a focus on AI misuse rather than AI accidents.

recommendations from the safety dialogue or its participants.[18] It may also be harder to attract participants to an event that is seen as politically biased.[19]

There are several ways in which organizers can ensure that safety dialogues are non-partisan in the relevant ways, and that they are seen as such:

- Use non-partisan funding (such as from non-partisan foundations) to pay for the safety dialogue. In this context, "non-partisan foundations" are primarily ones that are not strongly associated with a particular political party, or with a particularly hawkish or dovish stance towards other countries that will be represented at the relevant safety dialogue.[20]
- Invite participants with diverse views about the other countries that are sending participants. This would reduce the likelihood of participants being exclusively hawkish or dovish towards the other countries.
- Safety dialogues should have a spirit of collaborative truth-seeking, even when participants have ideological disagreements, as discussed immediately below.

Promote a spirit of collaborative truth-seeking

We hope that participants will approach safety dialogues with a collaborative and truth-seeking mindset. This is in contrast to a mindset that sees the dialogues as more adversarial, or like a negotiation. There is some reason to think that participants would have a more adversarial mindset, unless organizers work to avoid this; AI development, especially in relation to the US and China, is often seen through a competitive and zero-sum lens (Toner et al., 2023; Zwetsloot et al., 2018).[21]

A collaborative truth-seeking mindset might have two benefits over a more adversarial mindset. First, this mindset would help participants to arrive at beliefs about AI risks that are more likely to be true, better informing any efforts to reduce these risks. Second, this mindset would promote cooperative relationships between participants, making them better able to work together, including after a safety dialogue, to reduce risks.

[18] An example – though one that highlights the difficulty of being seen as neutral – was that the Pugwash Conferences were seen by different actors as both too pro- and too anti-Communist. Groups avoided engaging with ideas from Pugwash for both of these reasons (Barrett, 2020, p. 194; Kraft & Sachse, 2020, p. 3).

[19] For example, participants may disagree with the bias, be concerned about the reputational costs of attending, or be blocked from attending by the organizations with which they are affiliated.

[20] The Pugwash Conferences seem to have lost credibility in the US because they received a lot of funding from Cyrus Eaton, who was seen as biased towards communism (Rubinson, 2020, p. 166). In a track II context, Jones (2015, pp. 199–205) discusses in detail the importance of non-partisan funding.

[21] More generally, the overall US-China relationship is often seen as zero-sum (Weiss, 2022).

This mindset seems to have been helpful in past cases. In the Pugwash Conferences, for example, organizers conceptualized these events as primarily "scientific." A key part of this vision was the belief that participants, as scientists, were particularly able to cooperate, despite the hostility between their home countries (Kraft & Sachse, 2020, pp. 13–16).[22] Reality did not always match this vision, but Kraft and Sachse (2020, p. 16) write that the cooperative atmosphere that it promoted was nevertheless somewhat helpful for Pugwash's goals.[23] Additionally, Jones identifies the "Problem-Solving Workshop" as a best practice in track II processes. Similar to the collaborative truth-seeking mindset, "these discussions are not meant to be forums where positions are repeated, but rather where joint analysis can lead to agreed-on understandings of the underlying causes of the dispute" (Jones, 2015, p. 114).[24]

There are several ways in which organizers can contribute to a spirit of collaborative truth-seeking during safety dialogues:

- Demonstrate collaborative truth-seeking. For example, people who are facilitating a safety dialogue can: ask open-ended questions to understand participants' views rather than interrogating their beliefs; restate participants' views to check understanding before responding; highlight areas of agreement between participants before exploring disagreements; seek to synthesize different perspectives into new solutions, rather than treating disagreements as zero-sum conflicts.
- Encourage participants to identify shared goals related to AI safety at the start of the dialogue. Highlighting these shared goals could help participants to temporarily set aside broader differences and stay focused on reducing risks. Similarly, facilitators could explicitly ask participants to set aside broader disagreements that they may have, in order to more productively work together to reduce shared risks from AI.[25]

[22] See, for example, page 15: "A second Pugwash claim emphasized that scientists as scientists were able to suspend national, political and ideological allegiances – at least temporarily – and that this afforded a means to transcend the ideological and political divides."
[23] "Around the Pugwash table, national allegiances and ideological affinities proved impossible to relinquish [...] Here, we see the myths coming centrally into play. In encouraging scientists to look to each other across the bloc divide, they helped to foster a sense of community and of loyalty to something other than the nation state – even if this was contingent, ephemeral and unstable. This perhaps helped to maintain levels of goodwill between scientists that could keep alive their commitment to the Pugwash project during periods of rancor and hostility."
[24] Jones continues: "[This joint analysis is] followed by joint development of ideas and options that would not be apparent from traditional zero-sum bargaining."
[25] The particular case that we have in mind is that AI safety discussions between participants from US and Chinese institutions may be more productive if these participants can set aside broader disagreements, e.g., about the role of the US and China, or about the relationship between these countries. During safety dialogues, we expect that it often will indeed be worth setting these questions aside in order to promote AI safety cooperation. That said, this is a difficult trade-off because these broader questions about the US and China are of course important.

- Encourage technical discussions between participants, such as by including these discussions in the agenda.[26] Relevant agenda items could include discussions of research findings about AI alignment challenges, proposals for safer AI development practices, and brainstorming new technical solutions to AI risks. We expect that it would be easier in technical discussions for participants to have the type of mindset that we describe.[27] That said, broader policy dialogues are also important, so organizers should think carefully about how to balance technical and policy discussions.

Create high-trust relationships between the participants

If participants have a high-trust relationship with each other, they might be better able to have productive conversations during the safety dialogue. Additionally, they might be better able to coordinate after the safety dialogue to get desirable AI safety interventions implemented.

Here are some examples of how organizers could contribute to these high-trust relationships:

- Encourage social interactions between participants. Shared activities and informal interactions might help participants to develop stronger relationships with each other, promoting trust, and thus making progress easier. This phenomenon seems to have been helpful both for the Pugwash Conferences, as well as for various track II dialogues (Jones, 2015, pp. 127–128; Kraft & Sachse, 2020, p. 22).
- Ensure that participants understand the limits of each other’s influence. Different participants will inevitably have differing levels of influence over the organizations with which they are affiliated.[28] Participants should understand each other’s level of influence, or at least that other participants’ influence may be low. If not, trust and interpersonal relationships might be damaged when, for example, an organization does

[26] The case of the Asilomar Conference suggests an additional benefit to promoting these kinds of discussions. Some of the discussions at Asilomar focused on technical ways to reduce the risk from recombinant DNA, such as how to create modified organisms that would be unable to live outside the lab. Paul Berg, the main Asilomar organizer, thinks that the participants (who were mostly scientists) found these discussions particularly interesting, making them more engaged during the conference in general (Grace, 2015, p. 26).

[27] For example, technical problems often have more clear "right answers" that people can work towards together, whereas political issues tend to be more intractable disagreements. Working together on technical issues can build relationships and habits of collaboration that make later policy dialogues more constructive.

[28] For example, due to differing seniority levels or differences in the ease of influencing different institutions.

something that is inconsistent with the views expressed by participants associated with that organization.[29]

- Promote a spirit of collaborative truth-seeking, as discussed above. As well as the benefits described previously, this ethos might contribute to high-trust relationships because it gives the participants the experience of collaborating with each other.[30]

Create high-trust relationships between the participants and facilitators

The literature on track II diplomacy stresses the importance of a high-trust relationship between people who are facilitating the process and participants, and notes that this relationship might not exist by default (Burgess & Burgess, 2010, p. 50; Jones, 2015, pp. 102–104).[31] Those authors suggest various ways in which facilitators can contribute to a high-trust relationship. Two ways in particular might apply well to safety dialogues:[32]

- Facilitators should demonstrate that they are listening to participants and care about their perspectives. For example, they can ask nonthreatening and open-ended questions to avoid appearing judgmental.
- Facilitators should be open about their motivations for taking the role; if these motivations come out later and are not what participants expect, this is harmful for trust.[33]

Communicating about safety dialogues to outsiders

Maintain confidentiality about what was said by whom

We recommend that organizers implement a "Chatham House" norm where participants can use the information that they learn from a safety dialogue, but not reveal the identity of the speaker (Chatham House Rule, n.d.). This norm would make it easier for participants to speak freely, contributing to

[29] An additional example: By default, one participant not committing to changing their organization's behavior may erode goodwill among other participants, but this may be mitigated if other participants understand that this lack of commitment might stem from the participant's insufficient influence to alter their organization's behavior.
[30] This is in contrast to, for example, a negotiation, which we expect would feel more adversarial and thus less good for creating trust.
[31] In our terminology, "facilitators" is a subset of "organizers" and refers specifically to people who help participants to have more productive conversations with each other, such as by chairing discussions. Although this best practice seems to apply in particular to facilitators, it might also be helpful for organizers more generally to exhibit these behaviors.
[32] Note, however, that these sources are particularly referring to track II events that aim to resolve violent conflict. We are unsure to what extent this best practice would be relevant in this case.
[33] Relatedly, Jones (2015, p. 78) recommends being impartial but not neutral. This means treating the participants fairly but expressing a viewpoint if one does have a strong view.

high-quality discussion.[34] Additionally, this norm might prevent a beneficial proposal for increasing AI safety from becoming controversial because it is known to have originated from a particular source.[35] That said, confidentiality may reduce accountability in significant and harmful ways. We expect that confidentiality is overall beneficial for the specific safety dialogue that we have in mind, but that it may do more harm than good in other types of safety dialogue, such as intergovernmental meetings.

Consider maintaining confidentiality about who is attending

It might be beneficial for there to be confidentiality from the public about who is attending, or at least for only some attendees' identities to be made public.[36] Organizers should consider setting this norm, but we are unsure whether organizers should in fact implement it; we expect that the answer will vary from case to case.

Keeping the attendee list of safety dialogues confidential from the public might be beneficial for at least two reasons. First, it could lower political risks for participants, encouraging more attendance. Second, it could prevent the event's participants or discussion topics from being associated with controversial figures who attend. This reduces the likelihood that participants or discussion topics would lose credibility among non-participants because of their associations.

On the other hand, there is also at least one reason it might be counterproductive to keep participants' identities confidential from the public: This might prevent participants from gaining credibility from being seen to be working to increase safety. That credibility could provide a further incentive to attend and increase the participants' ability to actually improve AI safety.

[34] The Pugwash organizing committee saw this dynamic as essential to the effectiveness of those conferences (Kraft & Sachse, 2020, pp. 21–22).

[35] On the other hand, Paul Berg, the main organizer of the Asilomar Conference, thinks that inviting journalists to attend the conference increased the legitimacy of its resulting recommendations; it protected against the conference being seen as "a secretive meeting of scientists, coming out with some conclusion that everybody had to live with" (Grace, 2015, p. 25). We suspect that this consideration is generally less important than the opposing considerations highlighted in the main text. That said, if widespread legitimacy is particularly important for a particular safety dialogue, organizers should consider taking the Asilomar approach.

[36] Note that we do not think that it would be possible or desirable for organizers to keep the identities of participants confidential from governments. Even so, confidentiality from the public could still be helpful. This applies even if the goal of a given safety dialogue is to improve government policy; public backlash against an individual because they attended a safety dialogue could make governments less willing to engage with that individual, even if the government knows in either case about the attendance. As an example, the Johnson Administration avoided engagement with Pugwash for fear of political attacks that it was interacting with communist stooges (Rubinson, 2019, p. 13).

It might therefore be best to make public the identities of attendees for whom the benefits of that outweigh the costs, but not of other attendees, and to note publicly that this partial confidentiality approach is being taken.

Consider publishing a readout after the dialogue

We expect that much of the influence of the safety dialogue described in this report would come from participants interacting privately with the organizations with which they are associated. That said, there could also be several advantages to publishing a public readout at the end of the safety dialogue:

- The readout could highlight participants' concerns about AI safety, contributing to (common) knowledge that some experts are concerned about risks from advanced AI. See joint statements organized by the Future of Life Institute and the Center for AI Safety for earlier examples of efforts to create this knowledge.
- If a safety dialogue includes participants from institutions in the US and China, then the readout would contribute to knowledge that experts from both the US and China are concerned about AI safety risks.[37] We hope that this would encourage efforts for these countries to put aside their disagreements in other areas in order to cooperate to reduce shared risks from AI. This might also help avoid a scenario where US and Chinese actors both avoid implementing AI safety measures because of a belief that these will not be reciprocated.[38]
- The readout could direct people who are concerned about AI safety to specific next steps that would be helpful. For example, the readout could suggest policies or interventions that the participants believe would reduce risks and would like to see implemented.[39] Additionally, the readout could highlight topics where the participants would like more research because they are still uncertain or without consensus. This could be a good way to direct talent and funding towards important research questions.

We suggested above that it might be better for the identities of participants not to be revealed publicly. If organizers do choose this approach, then it might be harder to publish a readout that attracts significant attention. That said, the

[37] Note that some of the signatories of the FLI and CAIS statements were experts in China.
[38] See Toner et al. (2023) for some early examples of this scenario playing out.
[39] Indeed, recommendations in the Asilomar summary statement were adopted by the US National Institutes of Health, as well as by similar organizations in other countries (Grace, 2015, p. 1).

readout could presumably at least indicate the kinds of people who attended, or name some of the participants, and so attract attention in these ways.

Content of the event

Provide inputs to encourage participants down a productive path

Jones (2015, pp. 105–108) describes three "inputs" that facilitators can provide to help participants in track II processes.[40] We expect that this taxonomy would also help people who are facilitating safety dialogues to conceptualize their contributions to safety dialogue conversations.[41] We present a slightly adapted version of it:

- Theoretical inputs are comments that introduce existing concepts, models, and empirical findings.[42] For example, facilitators could provide information about technical AI safety findings. Similarly, facilitators could inform participants about possible interventions identified by the AI governance field, such as the interventions described in section four of this report.
- Content observations are interpretations of the content that is being said by participants. This can help participants to understand each other better, such as if some participants do not articulate their views clearly, or if participants talk past each other, such as because of cultural differences.
- Process observations highlight how the event is unfolding and the broader implications of this. Jones gives an example where two groups of participants each believed that the politics of their own side was complex and messy, while the politics of the other side was ordered. By highlighting these contradictory perceptions, Jones broke a deadlock that stemmed from each side thinking that the other had to be the first to initiate action.

Sometimes split participants into working groups

If participants are struggling to make progress on a particular topic, or are "deadlocked," one solution might be to delegate the question to a smaller "working group." Working groups can make it easier to create consensus in multiple ways (Burgess & Burgess, 2010, p. 58; Jones, 2015, p. 72):

[40] The taxonomy was initially articulated by Kelmen and Cohen, though Jones adds various examples and explanations from his own experience of track II processes.
[41] By "facilitating," we mean that the person is helping participants to have more productive conversations with each other, such as by chairing a discussion.
[42] Jones is referring specifically to findings from conflict research, but this detail does not apply so well to the safety dialogue case.

- Meaningful discussions between smaller groups of people are often easier.
- Organizers can select participants who have relevant expertise to be in the working group. This might be helpful if the topic at hand requires specialized knowledge.
- Organizers can select participants who will find it particularly easy to have productive conversations with each other.[43]

Selecting participants to invite

Choose participants who will engage constructively

We recommend that organizers select participants who are likely to engage constructively with the process. Experts on track II diplomacy have identified several characteristics that lead to constructive engagement (Burgess & Burgess, 2010, p. 41; Jones, 2015, pp. 123–126). We suspect that many of these would also apply to safety dialogues. We list them here, adapting them slightly to this context. Participants should:

- See the problem realistically and appreciate how difficult it might be to solve. In this context, "the problem" could be both the technical difficulties around AI safety, as well as the political difficulties that might make it harder to reduce AI risks, such as international tensions.
- Have a tendency to follow the rules and norms that are established for the safety dialogue.
- Be well-connected to groups that the safety dialogue would ideally influence.
- Be ready to look beyond official positions and to develop alternative and new ideas.
- Want to solve the issues that are being discussed.
- Have good interpersonal skills.
- Be seen as thoughtful, honest, and trustworthy.

Consider including participants from a range of countries

This report focuses particularly on dialogues involving participants from at least US and Chinese institutions. That said, even if organizers are primarily interested in creating relationships between these countries, they should

[43] As an example from track II diplomacy, Jones (2015, p. 73) notes that a working group composed of Indian and Pakistani submariners might break through a deadlock in the broader Indian-Pakistani group; the shared experience of working on submarines creates a stronger basis for communication and trust.

consider including participants from a wider range of countries, for several reasons:

- AI expertise is not just concentrated in the US and China, and various other jurisdictions are also important for AI governance (Maas, 2023, pp. 89–91).
- Forums that are only attended by a small number of great powers are sometimes seen as less legitimate by key stakeholders.[44]
- Including third countries might make safety dialogue participants less likely to view the event through the frame of zero-sum competition between the US and China.[45] Avoiding this frame might make it easier for participants from the US and China to find areas where they or their institutions can cooperate to promote safety, even if the overall relationship between their countries is poor.

We expect it would generally be better to include participants from a range of countries. That said, there are some disadvantages compared to a dialogue between participants from just the US and China:

- It is often harder for a wider range of actors to reach consensus or agreement; they are likely to have a wider range of views and incentives.
- Many third countries that have leading AI experts are US allies.[46] As a result, including third country experts might make the safety dialogue feel skewed against China. As an alternative, organizers could deliberately invite participants from neutral or China-allied countries. This might mean, however, that there are more participants without highly relevant backgrounds, potentially reducing the quality of discussion or the potential influence of the safety dialogue.

Consider the right level of participant "turnover" between dialogues

If there are multiple dialogues over time, organizers would need to decide how much participant "turnover" they would want between these different events, i.e., the extent to which it is the same people attending each event. Organizers should try to strike a balance between competing considerations. On the one hand, a high turnover rate would increase the number of participants that meet

[44] There is some discussion of this point in relation to AI governance in Cihon et al. (2020, p. 553) and Ho et al. (2023, pp. 8–9).

[45] The US-China relationship, including in relation to AI, is often framed in terms of zero-sum competition (Toner et al., 2023; Weiss, 2022; Zwetsloot et al., 2018).

[46] For example, out of the ten countries that have produced the most AI research (as counted by published papers or citations), only one (India) is not either the US, China, or a clear US ally (Chahal et al., 2022).

each other, potentially creating more relationships.[47] On the other hand, repeated interactions between the same participants could be helpful for creating particularly strong relationships (Burgess & Burgess, 2010, pp. 58–59). Additionally, if participants have already attended a dialogue, they are more likely to be familiar with the norms for how the dialogue is run.

Logistical details

Choose a suitable location

Based on his experience of running track II events, Jones (2015, pp. 126–128) identifies several important characteristics for choosing a location for track II events. We expect that these characteristics would generally also be valuable for safety dialogues:

- Social opportunities: The location should allow for shared experiences such as meals and excursions. These would promote informal interactions and relationship-building among participants.[48]
- Accessibility: Choose a location that is easy for participants to reach. This involves not only geographical considerations but also logistical factors such as air connections and visa requirements.[49]
- Comfort: The venue should be comfortable so that participants are in a good mood and well-rested, as well as to show appreciation for participants attending.
- Neutral ground: The venue should ideally be located on neutral ground: a place that is not perceived as providing any side with more symbolic control. This might make participants more able or willing to attend and make them more comfortable speaking freely. For example, Singapore and Switzerland each have fairly good relations with both the US and China (Chong, 2023; Grano & Weber, 2023).

[47] There may be other benefits to higher levels of turnover. For example, if new types of expertise are needed, a higher turnover level will make it easier to select participants who have the necessary expertise (Jones, 2015, pp. 125–126).

[48] This approach was adopted at the Pugwash Conferences, where "a busy social program [...] enhanced further the scope for informal conversations" (Kraft & Sachse, 2020, p. 22).

[49] For example, potential Pugwash participants were sometimes unable to attend because of visa issues (Rubinson, 2020, p. 163).

Reduce language barriers

AI experts, including in China, often have good knowledge of English.[50] That said, there may still be some cultural or language barriers among safety dialogue participants.[51] The topics of safety dialogues are likely to be extremely technical. Additionally, some AI safety terms have different nuances in English and Chinese, making it easier for people who use these terms to talk past each other (Imbrie & Kania, 2019, pp. 4–5).[52]

There are various ways to reduce these barriers:

- Organizers could select partly for particularly fluent English when choosing participants to invite. This might have the drawback of excluding some otherwise promising participants.
- The safety dialogue could include people who are attending as translators rather than as participants.[53]
- Experts could convene to create glossaries explaining the nuances of technical terms in different languages.[54] The process of creating these glossaries could itself be helpful for promoting international cooperation on AI safety (Imbrie & Kania, 2019, pp. 4–5).[55]

[50] We base this claim on the description in ÓhÉigeartaigh et al. (2020, p. 579) of organizing a cross-cultural AI workshop, the fact that the large AI conferences listed in Maslej et al. (2023, p. 65) all take place in English, and the general trend of English being a lingua franca. That said, we expect that there are some talented individuals in the field who do not have advanced English skills (and who might as a result be shut out of events such as conferences).

[51] We focus here on barriers between speakers of Chinese and English. That said, we expect that similar points would apply to some extent whenever a safety dialogue involves speakers of different languages.

[52] Relatedly, subtly incorrect translations of key terms in Chinese documents seem to have contributed to misunderstandings in the US about Chinese AI policy (ÓhÉigeartaigh et al., 2020, p. 587).

[53] This option does have some drawbacks compared to an ideal world where participants can speak freely with each other without translators. For example, waiting for a translation might add friction to a conversation. To make translation work as well as possible, we recommend that organizers look for translators who do not just have good language skills but also relevant subject expertise such as knowledge of machine learning.

[54] Glossaries could be made by the participants of a particular safety dialogue or by a separate group of experts.

[55] Wheeler (2014, pp. 28–29) notes that making a shared glossary seems to have been helpful for promoting US-China cooperation in the realm of nuclear security, and provides some insights that may be helpful if organizing an AI safety glossary.

3. Harmful outcomes to avoid

International AI safety dialogues could lead to harmful outcomes, reducing the overall value of the safety dialogue, or even causing the safety dialogue to do more harm than good. We sketch out possible harmful outcomes that are particularly concerning to us and suggest ways for organizers to reduce these risks.

Although organizers should try to reduce the risks that we describe, it will often be good for a safety dialogue to go ahead, even if the risks cannot be completely eliminated; the relevant question is not whether harm might occur, but whether the potential benefits outweigh the risks.

Promoting interest in AI capabilities disproportionately, relative to AI safety

The problem

Discussions of AI safety issues might increase people's beliefs that future AI systems could be extremely powerful; arguments that AI could be capable of causing such extreme catastrophes generally involve the claim that AI could be very capable. This could backfire if people in these discussions become more concerned about ensuring that their group has such powerful systems rather than on ensuring that powerful AI systems are safe; as a result, these people might promote progress on making AI more capable excessively in relation to safety progress.[56] We expect that promoting the rapid development of AI capabilities without sufficient emphasis on safety would increase AI safety risks. As a result, it would be harmful for safety dialogues to make participants internalize ideas about AI power but not ideas about AI risks.

There are, of course, benefits to many AI developments, and there are important strategic considerations to consider about who develops particular kinds of systems. We expect, however, that accident risks will tend to receive insufficient attention relative to these potential benefits and strategic implications. There are strong incentives to pursue beneficial AI deployments. It may also be easier for people to grasp some possible use cases of powerful AI than the possible risks; the arguments for AI risks can sometimes seem speculative. Additionally, participants' backgrounds might affect what ideas they

[56] Anecdotally, we and various people we know in the policy and AI research communities have had repeated experiences where others clearly reacted to information about large-scale AI risks partly by (at least temporarily) becoming more interested in and believing in the potential potency of AI, and being more interested in (their groups) developing or deploying it faster.

internalize. For example, participants who are directly involved in AI development might be naturally inclined to focus on AI's potential.

Possible Solutions

We encourage facilitators to keep participants on topic, i.e., safety concerns around AI. Additionally, it may be helpful to not focus discussions much on the potential military applications of AI, such as for cyberattacks or detecting submarines (Maas et al., 2022). National security is often framed in a zero-sum way, and so is particularly likely to promote thinking about competitive strategic dynamics, rather than accident risks that might affect everyone.[57]

Reducing the influence of safety concerns

The problem

Poorly managed safety dialogues might reduce the influence over AI development of safety-focused actors. These actors might include specific individuals (such as safety dialogue participants), as well as the AI safety community more broadly. We expect that reducing the influence of safety-focused individuals and institutions would reduce the likelihood that advanced AI is developed and deployed in safe ways.

One example of safety dialogues reducing influence could happen if there are participants from both the United States and China; participants from these countries might be perceived as engaging with geopolitical adversaries. This could be seen as illegitimate or treacherous by some observers, possibly reducing their influence.[58] As a different example, government officials may interpret the actions of participants as attempts to conduct foreign policy on their behalf. This might cause a backlash, further reducing the influence of participants or the field in general. Additionally, a safety dialogue itself might become seen as tainted, reducing the influence of the dialogue as an institution. This seems to have happened in historical cases, such as sometimes with the Pugwash Conferences (Rubinson, 2019, pp. 156–158).

Possible solutions

There are strong arguments for the value of safety dialogues. Giving these arguments should often be sufficient to rebut any criticism of safety dialogues as a concept, or of the people who attend them. In particular, because AI catastrophes could have extremely wide-ranging effects, everyone benefits from

[57] See, for example, discussion of the "security dilemma" (Rittberger, 2004, pp. 3–4).

[58] As an example, various figures associated with the Pugwash Conferences suffered this effect (Rubinson, 2020, pp. 156–158).

AI-relevant actors having a better understanding of safety concerns; this remains the case even if those actors are rivals or are based in countries that one does not like. Indeed, there is ample precedent of the US and China, as well as other geopolitical rivals, having dialogues or otherwise communicating in order to reduce risks that affect them both (Haenle, 2021; Imbrie & Kania, 2019, pp. 6–7; Miller, 2021).

As well as making these arguments, organizers can take additional steps to reduce the likelihood of safety dialogues reducing the influence of safety-focused actors:

- Frame the event as not attempting to benefit some countries over others. This would reduce the chance of the dialogue being misinterpreted as a political maneuver, and reaffirm its purpose as a venue for addressing important safety issues. It would also underscore the idea that AI safety is of global concern, rather than confined to any particular country's agenda.
- Try to avoid conversations that might be perceived as making foreign policy without authorization. For example, facilitators can clarify that all participants are only representing themselves and keep the focus on AI-related topics rather than discussions of broader relationships between countries.[59]
- Invite participants from a range of countries so that the safety dialogue is less likely to be seen purely through the lens of US-China geopolitical competition. This would reduce issues around talking to geopolitical adversaries, such as the risk of it seeming illegitimate, or of appearing to be doing foreign policy. That said, there are trade-offs to inviting participants from a wider range of countries, as discussed above.
- Invite participants with diverse views about the other countries that are sending participants. This would reduce the likelihood of participants being exclusively hawkish or dovish towards the other countries, and of them being perceived as such. Such a perception might reduce the influence of the safety dialogues among audiences that are more or less hawkish than they perceive the dialogues to be.[60]

Diffusing AI capabilities insights

The problem

[59] International dialogues about sensitive topics often do this. See, for example, the Pugwash Conferences and track II diplomacy (Jones, 2015, p. 25; Kraft & Sachse, 2020, p. 15).

[60] We suggest elsewhere in the report that organizers consider not publicizing the identity of participants. If organizers do make this choice, then inviting participants with a range of political perspectives will presumably have less of an effect.

AI safety dialogues might spread insights about how to build more capable AI systems. We call this phenomenon "capabilities diffusion."[61] Capabilities diffusion might happen if the participants do technical AI work and want to discuss their work with each other, or if they are unable to have nuanced conversations about AI safety without discussing specific approaches in AI capabilities. The fact that AI safety insights and AI capabilities insights are so intertwined makes this phenomenon more likely to happen (Christiano et al., 2023, pp. 1–2; Hendrycks & Mazeika, 2022, pp. 7–9).

Capabilities diffusion could potentially be harmful in at least three ways.[62] First, institutions may be less willing to allow people associated with them to attend safety dialogues if they are worried that these people will diffuse capabilities to potential rivals. This might prevent people from participating who would otherwise participate and contribute to making the safety dialogue successful. Second, diffusion could speed up the rate of progress at the AI frontier. We expect that this increases large-scale AI risk by leaving less time for technical or governance work to reduce risks from AI systems before these systems are sufficiently advanced to potentially be extremely dangerous (Hendrycks et al., 2023, pp. 17–20; Hendrycks & Mazeika, 2022, pp. 7–9). That said, there is a complicated relationship between the pace of AI capabilities progress and safety.[63] Third, diffusion could broaden the number of actors who are able to build very advanced AI systems. This might be harmful via making coordination on safety measures around very advanced AI harder, and via increasing the likelihood of one actor recklessly or maliciously using advanced AI. That said, there are also legitimate reasons to want broader access to advanced AI, such as concerns about concentration of power (Cottier, 2022).

Capabilities diffusion should not, however, be overstated as a risk. The main reason for this is that the marginal effect of safety dialogues on diffusion seems small. If researchers want to share an insight about capabilities, they are already reasonably able to do so, such as by presenting at a conference or publishing a paper. If, by contrast, a participant already thinks that it is bad to diffuse capabilities insights, they would presumably try to stick to that principle also in the case of safety dialogues.[64] Indeed, if the safety dialogue includes participants from the US and China, participants might be particularly likely to be careful to

[61] Some sources, e.g., Anderljung et al. (2023, p. 14), instead use the term "proliferation."

[62] For some further discussion of capabilities diffusion, see Cottier (2022).

[63] Slower AI development could change several considerations in a way that might increase AI risk. For example, it might disproportionately affect safety-conscious actors, leave less time for safety work when that time could be most helpful, or increase multipolarity. For an accessible overview of relevant considerations, we recommend Stein-Perlman (2023).

[64] Participants might think this for safety reasons, but also for more pragmatic reasons. For example, they might not want to reveal information if the confidentiality of this information puts their company at a commercial advantage.

avoid sharing capabilities insights; the relationship between these countries is often seen in terms of tech competition and industrial espionage.[65]

Possible solutions

Organizers who are concerned about capabilities diffusion could take several steps to reduce its likelihood while still enabling valuable conversations about safety. They should establish clear guidelines against participants sharing specific technical insights that could accelerate AI capabilities progress, and vet potential participants to select those likely to respect such norms. Organizers could also try to structure discussions to focus as much as possible on conceptual issues related to safety rather than technical specifics of cutting-edge capabilities work, though disentangling the two may be challenging for detailed safety conversations. Additionally, they might want to consider limiting participation only to researchers who are not directly involved in advancing the capabilities frontier themselves, if the inclusion of cutting-edge capabilities researchers does not seem particularly valuable for that dialogue.

[65] That said, there might be capabilities diffusion in cases where participants are willing to discuss insights from their work, but where they have not (yet) done so in a conference or paper for some reason.

4. Interventions to discuss at safety dialogues

Discussing specific AI safety interventions at safety dialogues might be a helpful way to ground the discussion. It might also increase clarity about which interventions would be desirable, as well as technically and politically feasible. If a given intervention is desirable and feasible, safety dialogues could then contribute to implementing this intervention, such as by providing a space where details can be worked out and by helping relevant people to coordinate on implementing the intervention. Additionally, for any interventions that would require or benefit from international cooperation or coordination, safety dialogues might be a helpful "stepping stone" towards this.[66] For example, safety dialogues might improve trust between people of different countries and increase international consensus about the value of a given intervention.

We describe here several interventions that it might be fruitful to discuss in a safety dialogue, drawing on several prominent proposals from the AI governance and safety fields.[67] We do not mean to imply, however, that safety dialogue organizers or participants should assume that these interventions are desirable or feasible; reasonable people might disagree about this. Additionally, safety dialogues will generally work best if participants have a sense of "co-creating." As such, organizers should be sure to allow individual participants to meaningfully shape the results of the discussion, even if organizers present specific potential interventions.

This section is organized into three parts. We first summarize an overarching plan for avoiding large-scale risks from misalignment. Second, we highlight several best practices that could be implemented by AGI labs. Third, we do the same for best practices that could be implemented by other AI-relevant actors.

[66] International cooperation or coordination might be valuable in many cases to avoid "races to the bottom" on safety standards in different countries (Ho et al., 2023, pp. 5–6; Trager et al., 2023, p. 10).

[67] In particular, the "overarching plan" is similar to the measures proposed in two papers that were co-authored by many of the leading figures in those fields (Anderljung et al., 2023; Shevlane et al., 2023). Additionally, the best practices that we highlight scored highly in two recent surveys of those fields (Räuker & Aird, 2023; Schuett et al., 2023)

An overarching AI safety plan

In this subsection, we first sketch out a high-level plan for reducing large-scale risks from AI misalignment, drawing in particular on Anderljung et al. (2023) and Shevlane et al. (2023).

National-level licensing of cutting-edge training runs

AI labs could be required to obtain a license from a national regulator before training highly capable new models. The license could stipulate that the lab undertakes particular safety measures such as having pre-deployment safety evaluations carried out by a third party (see immediately below) and implementing strong cybersecurity to reduce the likelihood that powerful models are stolen by malicious or reckless actors. For more on national-level licensing, see Anderljung et al. (2023), in particular, section three.

Pre-deployment safety evaluations

As a condition for granting a license, policymakers could require AI labs to commit to safety evaluations of models, with models not being deployed if they do not pass these evaluations.[68] These evaluations could be carried out by external third parties to reduce undue incentives on the evaluator to declare a model safe.[69]

Shevlane et al. (2023) identify two kinds of safety evaluation: Dangerous capability evaluations assess the capability of a model to do harm, while alignment evaluations assess the propensity of a model to do harm. Dangerous capabilities are generally offensive capabilities; they are useful for gaining influence (e.g., manipulation) or threatening security (e.g., weapons acquisition). Some dangerous capabilities (e.g., self-proliferation) are capabilities that would be useful for a misaligned system that is attempting to evade human oversight.[70] Early work on safety evaluations is already underway, with ARC Evals doing dangerous capability evaluations on Claude and GPT-4 (Shevlane et al., 2023, p. 10).[71]

[68] Models can be dangerous before being fully trained. As a result, it might also be desirable to also have safety evaluations *partway through* training (Shevlane et al., 2023, p. 7). More speculatively, regulators could attempt to predict *before* training how dangerous the resulting model will be. For example, they could use scaling laws and information about the training run to attempt to predict the model's capabilities. See Appendix B of Shevlane et al. (2023) for more on scaling laws.

[69] See Anderljung et al. (2023, pp. 24–27) for additional detail on safety evaluations.

[70] For more detailed discussion of specific capabilities see Shevlane et al. (2023, pp. 4–5). Examples of model capabilities that would be useful for self-proliferation include breaking out of its local environment and independently generating revenue to pay for access to compute.

[71] Claude and GPT-4 are cutting-edge language models from Anthropic and OpenAI respectively. Dangerous capability evaluations were also performed on GPT-4 by groups other than ARC Evals.

Tracking large compute clusters

The training runs that produce advanced AI systems require large numbers of specialized AI-relevant chips ("compute"). Additionally, efficient training runs generally require the chips to be in close proximity to each other, i.e., in "clusters." These properties could provide a practical mechanism for enforcing regulations about AI development. Governments could set up systems to monitor who owns large amounts of compute – and large compute clusters in particular. For example, governments could track the possession of individual AI-relevant chips in registries. If an actor is known to have a lot of compute but is not applying for any training run licenses, then regulators could know to check for unauthorized training runs (Shavit, 2023, pp. 12–13; Whittlestone et al., 2022, 2023).[72]

Best practices for AI labs

In addition to the overarching plan discussed immediately above, it might be helpful for safety dialogues to discuss standalone interventions that could be implemented to potentially reduce risks. This subsection focuses on best practices that could be implemented by individual AI labs.[73]

Measures here are taken from Schuett et al. (2023). The authors surveyed 51 experts from labs, academia, and civil society about best practices for AGI labs. We cite the five measures that had the most support; for each measure here, 98% of respondents either "somewhat" or "strongly" agreed that AGI should implement it:[74]

1. **Pre-deployment risk assessment**. AGI labs should take extensive measures to identify, analyze, and evaluate risks from powerful models before deploying them.
2. **Dangerous capability evaluations**. AGI labs should run evaluations to assess their models' dangerous capabilities (e.g., misuse potential, ability to manipulate, and power-seeking behavior).
3. **Third-party model audits**. AGI labs should commission third-party model audits before deploying powerful models.

[72] One complication is that AI labs often rent compute from cloud compute providers, rather than buying it. This could be addressed with know-your-customer rules. Cloud compute providers could be obliged to only provide AI-relevant compute to actors with a training run license, or to notify regulators about the end user of this compute. See some discussion of this in Anderljung & Hazell (2023, p. 14).

[73] The following subsection focuses on best practices for other kinds of actors.

[74] Although the five measures here received the most agreement, all but one of the 50 measures in the survey had majority agreement. For additional measures that labs could implement, see Maas (2023, pp. 88–92) and the sources cited there.

4. **Safety restrictions**. AGI labs should establish appropriate 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).
5. **Red teaming**. AGI labs should commission external red teams before deploying powerful models.

As with the other measures highlighted in section four, discussing these measures at safety dialogues could be helpful for increasing international consensus about what interventions would be valuable. Two specific forms of increasing consensus seem particularly relevant in the context of these measures. First, it would be helpful for safety dialogues to explore whether Chinese experts would also be in favor of these measures, and whether the measures would also work well for Chinese labs; Schuett et al. (2023) focused in particular on labs in the West and seem to have only surveyed experts in the West. Second, participants at safety dialogues could attempt to find ways to make these measures more concrete, while still retaining a consensus that they are beneficial. The survey authors note that they described the measures in an abstract way and that this may have contributed to the high apparent consensus (Schuett et al., 2023, p. 10).

Best practices for other relevant actors

Actors other than AI labs could also implement best practices to increase AI safety, even without other actors reciprocating. We focus here on measures that were particularly popular in the Räuker and Aird (2023) survey and that could be done by actors in a heterogeneous group of countries (rather than, for example, being specific to the US context).[75]

- **Shift AI publication norms toward "don't always publish everything right away."** This might be beneficial because it would reduce capabilities diffusion – a problem discussed earlier in the report and in Cottier (2022).[76]
- **Improve the AI Incidents Database** by either building on the existing database or starting a better one. An improved database of safety failures might make it easier to communicate with policymakers and the general public about AI safety failures. It might also improve researchers' ability

[75] Whereas the Schuett et al. survey focuses specifically on what AGI labs should do, the Räuker and Aird survey has a broader scope; goals that experts think should receive more funding. This means that the Räuker and Aird survey can be used to identify best practices for a wider range of actors than just labs.

[76] That said, there are of course important disadvantages to shifting norms in these ways, such as slowing beneficial AI discoveries and going against values about open science. Decision-makers should carefully consider how to make the best trade-offs here.

to learn from these incidents. See McGregor (2020) for detail on the existing database.

- **Increase the liability of AI product providers for harms caused by their products**. This might incentivize progress on AI safety and security practices.

Acknowledgements

We are grateful to Ben Garfinkel, Seán Ó hÉigeartaigh, Olivia Shen, and the participants of an IAPS seminar for their discussion and input.[77] We are also grateful to Adam Papineau for copyediting.

[77] The views in this report do not necessarily reflect the views of these individuals.

Appendix: Additional detail on the "strand 1" case studies

Cases that we selected

We provide a little more detail here on the three past events that contribute to our recommendations of best practices.

Pugwash Conferences during the Cold War (from 1957)[78]

These conferences brought together scientists from both sides of the Iron Curtain to discuss disarmament, particularly in relation to nuclear weapons. The conferences are credited with reducing nuclear risk by contributing to the Limited Test Ban Treaty (1962) and the Anti-Ballistic Missile Treaty (1972) (Rubinson, 2019, p. 3). There are, however, limitations to what we can learn from this case. In particular, an example from decades ago might no longer have lessons for today. Additionally, in contrast to the kind of safety dialogue that we describe, Chinese participants had relatively little involvement with Pugwash (Barrett, 2020, p. 190). That said, there were many Soviet participants in Pugwash; the USSR during the Cold War may be more helpful than Cold War China as an example for thinking about contemporary China's relationship to the US.[79]

Asilomar Conference on Recombinant DNA Molecules (1975)

Asilomar brought together scientists to design influential safety guidelines for performing recombinant DNA research.[80] This is a good example of an event that reduced risks from new technology (Grace, 2015). That said, Asilomar has limited relevance in that it mostly did not bring together experts from countries with strained relationships; participants were mostly from institutions in the US or close allies, though there were a few from the USSR.[81]

[78] There continue to be Pugwash Conferences, but our impression is that these events had far more influence during the Cold War. Additionally, sources about Pugwash generally focus exclusively on the Cold War period. See, for example, Kraft & Sachse (2020) and Rubinson (2019).

[79] The current US-China relationship is (increasingly) framed in a similar way to the Cold War US-USSR relationship, e.g., in terms of great power competition, and even with the explicit framing of a "Second Cold War." See Imbrie & Kanie (2019, p. 6) for some discussion of the similarities and differences between these relationships.

[80] Recombinant DNA is a form of DNA created by combining genetic material from multiple sources. It could be accidentally or maliciously used to create harmful new organisms (Grace, 2015, pp. 3, 14).

[81] Asilomar participants are listed in the appendix of Fredrickson (1991). There were 86 participants from US institutions and 53 participants from non-US institutions. In our count, we

Cross-Cultural AI Ethics and Governance Workshop Series (from 2019)

These (approximately) yearly events explicitly bring together participants from Eastern and Western countries to improve international cooperation on AI, but without an explicit focus on catastrophic or existential risk (Center for Long-term Artificial Intelligence, 2022; ÓhÉigeartaigh et al., 2020).

Cases that we did not select

We identified some cases that are somewhat relevant to our criteria, but that are not sufficiently relevant to be included in this comparatively short report. We list these cases here in case they are helpful for more comprehensive work, and to give a better sense of our reasoning.

Specific track I, track 1.5, or track II events

Track II dialogues contribute to informing our "best practices"; they are the second knowledge "strand" that we draw on. That said, this strand mostly draws on best practices that have already been identified by practitioners in that field and that are not specific to safety dialogues. It might be helpful to look at specific track I, track 1.5, or track II dialogues and identify lessons that are relevant for safety dialogues in particular. Examples that might be particularly relevant include the US-China track 1.5 and track II dialogues about nuclear security (Wheeler, 2014), and the track I Shangri-La Dialogue, which involves both the US and China (Capie & Taylor, 2010).

Historical cases of US-China "exchange diplomacy"

These are many cases of non-governmental exchanges that potentially improved relations between participants' home countries – including in the US-China relationship. The most analogous exchanges to safety dialogues would be the ones between scientists.[82] That said, there are also well-known cases of exchanges between groups such as athletes and musicians (Millwood, 2022, p. 3).

Intergovernmental Panel on Climate Change

We avoided selecting this case because it involves far more participants than the event that we have in mind. Additionally, the IPCC is less focused on influencing

only found 12 participants from institutions in countries that were not at the time allied to the US: Poland (1), Switzerland (6), the USSR (5).

[82] Similarly, there may be lessons from science diplomacy. Science diplomacy consists of various practices at the intersection of science, technology, and foreign policy; it can involve both institutionalized roles, such as formal interactions between government officials, as well as non-institutionalized interactions, such as via informal channels (Melchor, 2020, pp. 411–412).

policy than safety dialogues might be, aiming to be “policy-relevant, but not policy-prescriptive” (Vardy et al., 2017, p. 56). That said, the attempt to reach international consensus about a major risk is clearly relevant to safety dialogues.[83]

[83] The IPCC has also been suggested as a possible model for a new international institution to improve the governance of advanced AI (Ho et al., 2023, p. 2).

Bibliography

Altman, S., Brockman, G., & Sutskever, I. (2023, May 22). Governance of superintelligence. OpenAI. https://perma.cc/DTG9-V5YX

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety (arXiv:1606.06565). arXiv. http://arxiv.org/abs/1606.06565

Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety (arXiv:2307.03718). arXiv. http://arxiv.org/abs/2307.03718

Anderljung, M., & Hazell, J. (2023). Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? (arXiv:2303.09377). arXiv. http://arxiv.org/abs/2303.09377

Barrett, G. (2020). Minding the Gap: Zhou Peiyuan, Dorothy Hodgkin, and the Durability of Sino-Pugwash Networks. In A. Kraft & C. Sachse (Eds.), Science, (Anti-)Communism and Diplomacy: The Pugwash Conferences on Science and World Affairs in the Early Cold War (pp. 190–217). BRILL. https://doi.org/10.1163/9789004340176

Bullock, J. B., Chen, Y.-C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. M., & Zhang, B. (Eds.). (2022). The Oxford Handbook of AI Governance. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.001.0001

Burgess, H., & Burgess, G. M. (2010). Conducting Track II peacemaking. United States Institute of Peace. https://perma.cc/4ZHD-NR3E

Capie, D., & Taylor, B. (2010). The Shangri-La Dialogue and the institutionalization of defence diplomacy in Asia. The Pacific Review, 23(3), 359–376. https://doi.org/10.1080/09512748.2010.481053

Center for Long-term Artificial Intelligence. (2022, December 10). Cross-Cultural AI Ethics and Governance Workshop Series. https://perma.cc/4SU2-KQVK

Chahal, H., Melot, J., Abdulla, S., Arnold, Z., & Rahkovsky, I. (2022, August). Country Activity Tracker. Center for Security and Emerging Technology; Center for Security and Emerging Technology. https://perma.cc/HD5K-VMQL

Chatham House Rule. (n.d.). Chatham House. Retrieved June 24, 2023, from https://perma.cc/MMR8-GZFC

Chong, J. I. (2023). Other Countries Are Small Countries, and That's Just a Fact: Singapore's Efforts to Navigate US–China Strategic Rivalry. In S. A. Grano & D. W. F. Huang (Eds.), China-US Competition (pp. 307–338). Springer International Publishing. https://doi.org/10.1007/978-3-031-15389-1_12

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2023). Deep reinforcement learning from human preferences (arXiv:1706.03741). arXiv. http://arxiv.org/abs/1706.03741

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the Future: Investigating Architectures for International AI Governance. Global Policy, 11(5), 545–556. https://doi.org/10.1111/1758-5899.12890

Cottier, B. (2022). Implications of large language model diffusion for AI governance (Understanding the Diffusion of Large Language Models). Rethink Priorities. https://perma.cc/HGS7-QHVX

Critch, A., & Krueger, D. (2020). AI Research Considerations for Human Existential Safety (ARCHES) (arXiv:2006.04948). arXiv. http://arxiv.org/abs/2006.04948

Dafoe, A. (2018). AI Governance: A Research Agenda. Centre for the Governance of AI. https://perma.cc/2GF6-Y7FV

Department for Science, Innovation and Technology. (2023, September 25). AI Safety Summit: Introduction. GOV.UK. https://perma.cc/X8LK-36BA

Ding, J., & Xiao, J. (2023). Recent Trends in China's Large Language Model Landscape. Centre for the Governance of AI. https://perma.cc/UH9P-4HNX

Fredrickson, D. S. (1991). Asilomar and Recombinant DNA: The End of the Beginning. In H. Ke (Ed.), Biomedical Politics. National Academies Press. https://perma.cc/8LV8-R4YN

Grace, K. (2015). The Asilomar Conference: A Case Study in Risk Mitigation. Machine Intelligence Research Institute. https://perma.cc/ALT7-L9VM

Grano, S. A., & Weber, R. (2023). Strategic Choices for Switzerland in the US-China Competition. In S. A. Grano & D. W. F. Huang (Eds.), China-US Competition (pp. 85–112). Springer International Publishing. https://doi.org/10.1007/978-3-031-15389-1_4

Haenle, P. (2021, August 11). Why the U.S. and Chinese Militaries Aren't Talking Much Anymore. Carnegie Endowment for International Peace. https://perma.cc/7T9S-VJDN

Heilmann, S. (Ed.). (2017). China's political system. Rowman & Littlefield.

Hendrycks, D., & Mazeika, M. (2022). X-Risk Analysis for AI Research (arXiv:2206.05862). arXiv. http://arxiv.org/abs/2206.05862

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks (arXiv:2306.12001). arXiv. http://arxiv.org/abs/2306.12001

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., & Snidal, D. (2023). International Institutions for Advanced AI (arXiv:2307.04699). arXiv. http://arxiv.org/abs/2307.04699

Hogarth, I. (2023, April 13). We must slow down the race to God-like AI. Financial Times. https://perma.cc/2WCS-PEQL

Imbrie, A., & Kania, E. (2019). AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement. Center for Security and Emerging Technology. https://doi.org/10.51593/20190051

Jones, P. L. (2015). Track two diplomacy in theory and practice. Stanford University Press.

Kraft, A., & Sachse, C. (2020). The Pugwash Conferences on Science and World Affairs: Vision, Rhetoric, Realities. In A. Kraft & C. Sachse (Eds.), Science, (Anti-)Communism and Diplomacy: The Pugwash Conferences on Science and World Affairs in the Early Cold War (pp. 1–39). BRILL. https://doi.org/10.1163/9789004340176

Lüscher, F. (2020). Party, Peers, Publicity: Overlapping Loyalties in Early Soviet Pugwash, 1955–1960. In A. Kraft & C. Sachse (Eds.), Science, (Anti-)Communism and Diplomacy: The Pugwash Conferences on Science and World Affairs in the Early Cold War (pp. 121–155). BRILL. https://doi.org/10.1163/9789004340176

Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. Contemporary Security Policy, 40(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464

Maas, M. M. (2023). Transformative AI Governance: A Literature Review [forthcoming] (4; AI Foundations Report). Legal Priorities Project.

Maas, M. M., Matteucci, K., & Cooke, D. (2022). Military Artificial Intelligence as Contributor to Global Catastrophic Risk. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4115010

Maham, P., & Küspert, S. (2023). Governing General Purpose AI — A Comprehensive Map of Unreliability, Misuse and Systemic Risks [Policy Brief]. Stiftung Neue Verantwortung. https://perma.cc/5DRF-QSXY

Marcus, G., & Reuel, A. (2023, April 18). The world needs an international agency for artificial intelligence, say two AI experts. The Economist. https://perma.cc/3PPG-Z4HA

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). The AI Index 2023 Annual Report. Institute for Human-Centered AI, Stanford University. https://perma.cc/9YSF-R5NS

McGregor, S. (2020). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database (arXiv:2011.08512). arXiv. http://arxiv.org/abs/2011.08512

Melchor, L. (2020). What Is a Science Diplomat? The Hague Journal of Diplomacy, 15(3), 409–423. https://doi.org/10.1163/1871191X-BJA10026

Miller, S. E. (2021). Nuclear Hotlines: Origins, Evolution, Applications. Journal for Peace and Nuclear Disarmament, 4, 176–191. https://doi.org/10.1080/25751654.2021.1903763

Millwood, P. (2022). Improbable diplomats: How ping-pong players, musicians, and scientists remade US-China relations. Cambridge University Press.

Ngo, R., Chan, L., & Mindermann, S. (2023). The alignment problem from a deep learning perspective (arXiv:2209.00626). arXiv. http://arxiv.org/abs/2209.00626

ÓhÉigeartaigh, S. S., Whittlestone, J., Liu, Y., Zeng, Y., & Liu, Z. (2020). Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. Philosophy & Technology, 33(4), 571–593. https://doi.org/10.1007/s13347-020-00402-x

Pause Giant AI Experiments: An Open Letter. (2023, March 22). Future of Life Institute. https://perma.cc/8RRU-M9PP

Perrigo, B. (2023, February 17). Bing's AI Is Threatening Users. That's No Laughing Matter. Time. https://perma.cc/L2E6-2ATM

Räuker, M., & Aird, M. (2023). Survey on intermediate goals in AI governance. Rethink Priorities. https://perma.cc/QXL3-HK5D

Rittberger, V. (2004). Approaches to the study of foreign policy derived from international relations theories. Institute for Political Science, University of Tübingen. https://perma.cc/HMM6-63TA

Rubinson, P. (2019). Pugwash Literature Review. Urban Institute. https://perma.cc/C8US-3S75

Rubinson, P. (2020). American Scientists in "Communist Conclaves:" Pugwash and Anti-communism in the United States, 1957–1968. In A. Kraft & C. Sachse (Eds.), Science, (Anti-)Communism and Diplomacy: The Pugwash Conferences on Science and World Affairs in the Early Cold War (pp. 156–189). BRILL. https://doi.org/10.1163/9789004340176

Russell, S. J. (2020). Human compatible: AI and the problem of control. Penguin Books.

Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023). Towards best practices in AGI safety and governance: A survey of expert opinion (arXiv:2305.07153). arXiv. http://arxiv.org/abs/2305.07153

Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring (arXiv:2303.11341). arXiv. https://doi.org/10.48550/arXiv.2303.11341

Sheehan, M. (2023). China's AI Regulations and How They Get Made. Carnegie Endowment for International Peace. https://perma.cc/5LBU-7PZ6

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). Model evaluation for extreme risks (arXiv:2305.15324). arXiv. http://arxiv.org/abs/2305.15324

Staats, J., Walsh, J., & Tucci, R. (2019, July 31). A Primer on Multi-track Diplomacy: How Does it Work? United States Institute of Peace. https://perma.cc/GGC2-ZNWJ

Statement on AI Risk. (2023, May 30). Center for AI Safety. https://perma.cc/FM27-GQ6E

Stein-Perlman, Z. (2023, April 17). Slowing AI: Foundations. LessWrong. https://perma.cc/2B6G-WHYB

Toner, H., Xiao, J., & Ding, J. (2023, June 2). The Illusion of China's AI Prowess. Foreign Affairs. https://perma.cc/2XDK-ZEF3

Trager, R. F., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., Ó hÉigeartaigh, S., Staffell, S., & Villalobos, J. J. (2023). International Governance of Civilian AI: A Jurisdictional Certification Approach. Oxford Martin AI Governance Initiative in partnership with the Centre for the Governance of AI. governance.ai/research-paper/international-governance-of-civilian-ai

Vallance, C. (2023, May 30). Artificial intelligence could lead to extinction, experts warn. BBC News. https://perma.cc/2ZGB-GAX9

Vardy, M., Oppenheimer, M., Dubash, N. K., O'Reilly, J., & Jamieson, D. (2017). The Intergovernmental Panel on Climate Change: Challenges and Opportunities. Annual Review of Environment and Resources, 42(1), 55–75. https://doi.org/10.1146/annurev-environ-102016-061053

Weiss, J. C. (2022, August 18). The China Trap. Foreign Affairs, 101(5). https://perma.cc/YN7K-LQRQ

Wheeler, M. O. (2014). Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned. Institute for Defense Analysis. https://perma.cc/T8MY-SPAV

Whittlestone, J., Avin, S., Collins, K., Clark, J., & Mueller, J. (2022, August 8). Future of compute review—Submission of evidence. CLTR. https://perma.cc/HSF2-ZZA3

Whittlestone, J., Avin, S., Heim, L., Anderljung, M., & Sastry, G. (2023, March 13). Response to the UK's Future of Compute Review: A missed opportunity to lead in compute governance. CLTR. https://perma.cc/LVV7-W97V

Zwetsloot, R., & Dafoe, A. (2019, February 11). Thinking About Risks From AI: Accidents, Misuse and Structure. Lawfare. https://perma.cc/8N7X-AD9M

Zwetsloot, R., Toner, H., & Ding, J. (2018, November 16). Beyond the AI Arms Race. Foreign Affairs. https://perma.cc/S5JQ-PTB9