DECEMBER 2019

AI Safety, Security, and Stability Among Great Powers

Options, Challenges, and Lessons Learned for Pragmatic Engagement

CSET Policy Brief

AUTHORS
Andrew Imbrie
Elsa B. Kania

Today's rapid advances in artificial intelligence and machine learning present a range of challenges and opportunities for the United States. Increasingly, U.S., Chinese, and Russian leaders recognize AI as a strategic technology that could become a critical determinant of future national competitiveness.[1] AI/ML may be poised to transform not only our economies and societies, but also the character of conflict.[2] The military applications of these technologies have generated particular concerns and exuberant expectations, including predictions that the advent of AI in military affairs could change the very nature of warfare.[3] Undeniably, AI has become a new focus of competition among great powers,[4] with the potential to disrupt the military balance and undermine deterrence.[5]

In this policy brief, we present and evaluate several measures in AI safety and security that could prove feasible and mutually beneficial for future bilateral and multilateral interactions. These measures are intended to prevent or correct misperceptions, enhance mutual transparency on policies and capabilities, and contribute to providing safeguards against inadvertent escalation. By pursuing such initiatives in the near term, the United States can improve its capacity to leverage the benefits of AI, while mitigating the risks and managing the shifting terrain in today's geopolitics, particularly among the United States, China, and Russia.

American strategy has reoriented toward great power rivalry, recognizing China and Russia as competitors that present a strategic challenge to the United States and its allies and partners worldwide.[6] This new direction demands creative thinking and solutions for complex policy issues. Any coherent framework for U.S. strategy must include policies to promote American innovation and competitiveness, while deepening coordination and collaboration with allies and partners. The reality of great power rivalry may entail sharper competition in areas where U.S. values and interests directly conflict with those of Beijing and Moscow, but it equally requires constructive approaches to pursuing selective, pragmatic, and carefully calibrated engagement on issues of mutual concern. In this new era, competition in AI technology and applications has emerged as a source of friction and created potential flashpoints. This initial assessment of AI safety and security concerns

illustrates one critical component of a long-term, comprehensive, and sustainable approach.

The dynamics of AI research are open and often collaborative, but the emerging discourse around AI has been growing increasingly fractured and competitive. The notion that an "AI arms race" is underway could exacerbate the challenges and misrepresent a range of emerging technologies that present complex and uncertain implications for the future of strategic stability.[7] Simply put, machine learning involves a set of interrelated techniques that can enable military capabilities, but these techniques do not themselves constitute weapons systems.[8] At times, military and political leaders have demonstrated more enthusiasm for AI/ML applications than awareness of the full range of risks and security concerns that could arise with the deployment of such nascent, relatively unproven technologies.[9] For instance, Russia is reportedly developing and planning to deploy by 2027 the "Poseidon," an underwater drone that will be armed with a nuclear warhead and capable of navigating autonomously.[10]

The challenges are acute and especially concerning, given the possibility of military powers rushing to deploy AI/ML-enabled systems that are unsafe, untested, or unreliable in an effort to gain a comparative advantage. Chief among the risks are failures, accidents, or unexpected emergent behaviors in AI systems that can exhibit unpredictable outcomes in real-world settings.[11] For military organizations, bureaucratic hurdles and the challenges of testing and assurance may slow adoption of these emerging capabilities, but the risks of accidents or adversarial interference cannot be discounted. Human-machine interactions will also create novel vectors of risk in the operation of highly automated or semi-autonomous systems.[12] AI systems also remain vulnerable to attacks, from the deliberate poisoning of data and cyber exploitation to the manipulation of brittleness or idiosyncrasies in algorithms.[13]

The rapid progress in dual-purpose research and applications in AI will heighten shared challenges to, and could worsen, relations among great powers. Following their deployment, interactions among AI systems could prove unpredictable in ways that intensify the risks of inadvertent escalation. Beyond the purview of nation-states, the diffusion of these technologies could empower non-state actors, from criminals to terrorist organizations, and present new security threats.[14] Most professional militaries are likely to operate in a manner generally consistent with the laws of war,[15] including the requirement for Article 36 review of new weapons systems to ensure their compliance with the Geneva Convention.[16] By contrast, non-state actors

could be uniquely empowered by the diffusion of emerging technologies—and unlikely to adhere to the same principles or parameters.

Given these concerns, there are compelling reasons to promote measures that enhance the safety, surety, and security of AI systems in military affairs. There are also difficult policy trade-offs involved. On the one hand, collaboration in AI safety and security can reduce the risks of accident and strategic miscalculations among great powers. On the other hand, such collaboration may improve the reliability of machine learning techniques and therefore enable strategic competitors to deploy AI/ML-enabled military systems more quickly and effectively. Evaluating the sensitivity of various countries to issues of safety, reliability, and assurance when fielding new weapons systems is beyond the scope of this paper, but merits further analytic attention. Nevertheless, any effort to promote collaboration in AI safety and security will need to balance the potential benefits against the range of possible costs.

Options and Recommendations for Pragmatic Engagement

American, Chinese, and international policymakers and stakeholders should pursue steps to improve transparency and promote mutual understanding of the factors that influence the design, development, and deployment of AI/ML techniques for military purposes. Over time, these measures could create a foundation for continued and collaborative initiatives to promote AI safety and security:

1. Develop common definitions and shared understanding of core concepts in AI safety and security and for AI in military affairs.

Among potential adversaries, and even allies and likeminded partners, differences in language and terminology can exacerbate misunderstandings. The field of AI today is relatively globalized, but there appear to be some discrepancies emerging in technical and doctrinal concepts.

As typically defined in the United States and Europe, the concept of "AI safety" refers to a cluster of research problems that deal with unintended harmful behavior or potential exploitation of machine learning systems.[17] AI safety constitutes a critical domain of research that is the subject of active inquiry and expanding activities within industry and academia worldwide. However, this research is often poorly understood and under-resourced, including in the United States, relative to the scope and scale of safety

concerns that may arise with the widespread deployment of AI systems.[18] U.S. policy initiatives to address issues of AI safety remain nascent and could encounter challenges in terms of incentives and implementation.[19] To date, Russia appears to be progressing more rapidly toward fielding and experimentation with unmanned, AI-enabled and potentially autonomous weapons systems, including in the course of its operational experiences in Syria.[20]

The Chinese government has started to promote research on and initial frameworks for AI safety. The Chinese approach to AI safety or security (人工智能安全, *rengong zhineng anquan*) appears to involve not only technical concerns but also questions about the impact on social stability and the security of the regime against potential threats to its authority.[21] This conceptualization is distinct from the issues under consideration by democratic governments.[22] Chinese concepts of AI safety and security have continued to evolve and progress, pursuant to active research and ongoing initiatives. For instance, the Beijing AI Principles (or "Beijing Consensus on AI," 人工智能北京共识), released in May 2019 by the Beijing Academy of Artificial Intelligence, included an emphasis on how to "control risks" through "continuous efforts ... to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys."[23] Notably, China's first "Guidelines on AI Safety/Security and Rule of/by Law" (人工智能安全与法治导则), which addressed issues of algorithm security, data security, intellectual property rights, societal employment (社会就业),[24] and legal responsibility, were released in August 2019.[25]

American, Chinese, Russian, and international policymakers should support the convening of technical experts from academia and industry to scope and define shared concepts, concerns, and research directions that involve the safety and robustness of AI systems. This effort should consider alternate framings, such as notions of reliable or robust AI.[26] Participants could elaborate definitions of the relevant terminology and shape constructive discourse, while identifying issues of mutual concern.[27] Such initial engagement, even on questions of definitions and aligning conceptual understanding, can facilitate productive dialogue and could become a basis for sharing best practices and devising sound policies to respond to common concerns.[28]

2. Pursue joint projects on a trial basis to summarize and evaluate each other's literature on AI safety and other relevant topics, while promoting transparency in AI safety and security research.

Even if trust is lacking and tensions are high, collaboration on carefully chosen initiatives can create a foundation for improved understanding and transparency. For instance, the United States and China have a long history of and continued engagement in scientific and technological collaboration that includes global health issues, such as capacity building on influenza surveillance.[29] This collaboration has often proved mutually beneficial,[30] despite ongoing concerns that such exchanges could be exploited for technology transfer.[31] There are also notable examples of productive bilateral or multilateral cooperation on nuclear issues, including cooperation to facilitate the removal and security of highly enriched uranium from a research reactor in Nigeria to mitigate the risks of non-proliferation.[32]

Even as the great power competition extends to new functional and geographic domains, there are reasons to sustain carefully calibrated engagement with competitors. Current dynamics among the United States, China, and Russia are different from the Cold War, but certain antecedents in Cold War history may hold lessons for today's challenges. Against the backdrop of mistrust and rivalry between the United States and Soviet Union, the Apollo-Soyuz Test Project brought together American and Soviet scientists and engineers for the first international human space flight, involving a rendezvous and joint experiments.[33] Despite persistent geopolitical tensions, scientists from the United States and Soviet Union continued to pursue select scientific and research exchanges, such as biomedical cooperation.[34] These exchanges provided an opportunity for improved understanding, including greater mutual visibility on technological advancements. Today, major powers should consider ways to build on those historical precedents.

Sino-U.S. relations are far more complex and economically interdependent than U.S.-Soviet relations were during the Cold War, including extensive co-authorship and collaboration in AI research.[35] Although a tendency toward strategic distrust on both sides may remain an enduring feature of the U.S.-China relationship, pragmatic engagement on discrete issues could serve the interests of both sides. In AI safety, an initial joint project could include collaboration between Chinese- and English-speaking countries to translate, summarize, and evaluate each other's scientific literature on AI safety and security. These efforts could focus on promoting transparency into applications of AI safety research, such as joint projects on verifying and validating systems for autonomous vehicles. One mechanism for future

collaboration could be the U.S. National Science Foundation and the National Natural Science Foundation of China program co-funding research undertaken by joint U.S.-China teams.[36]

3. Facilitate further Track 2 and Track 1.5 dialogues on concrete problems in AI safety and related security concerns.

The United States should promote and facilitate academic exchanges and research collaborations that bring together technical experts and social scientists from American, Chinese, Russian, and international institutions. Such dialogues could discuss concrete issues in AI safety, such as reward hacking, robustness to shifts in context, scalable oversight, verification and validation protocols, and progress toward reliable and interpretable AI/ML systems.[37] Industry stakeholders and policymakers could support these dialogues by providing resources and convening working groups on the sidelines of diplomatic engagements and international conferences.

These expert working groups could focus on specific topics in AI safety and then share their findings publicly or with their governments through designated channels. As with the Pugwash conferences during the Cold War, these dialogues could facilitate mutual understanding among scientific and technical experts.[38] Dialogues on AI safety could be sustained with appropriate parameters and assessments of risk in place, despite contentious geopolitical circumstances.[39] The international landscape for research and innovation is far more globalized than during the Cold War, and a more multilateral approach could prove effective in defusing and mitigating bilateral tensions and introducing a greater diversity of perspectives.

At the same time, gaps often exist in outlook between scientists and policymakers, such that trust and potential outcomes from Track 2 initiatives cannot be expected to translate directly or necessarily into Track 1 progress. Working groups and conferences could gradually incorporate participation from government officials to elevate these discussions to the Track 1.5 level.[40] Alternatively, participants could develop tighter feedback mechanisms between Track 2 conversations among non-governmental experts in order to inform future Track 1 official dialogues among government officials.

4. Develop common standards and shared methodologies of testing, evaluating, verifying, and validating systems for AI products and systems, including in such global industries as healthcare and autonomous driving.

U.S., Chinese, and international policymakers could support the convening of a group of experts from academia and industry to discuss common safety, testing, and validation standards around self-driving cars and other AI products, such as medical devices. This effort could include enlisting officials from the National Institute of Standards and Technology and its counterparts from Chinese, European, and other relevant institutions to discuss shared standards for testing autonomous vehicles and the vulnerabilities of these systems to adversarial attacks and systemic failures. Policymakers could support and leverage existing initiatives through the Institute of Electrical and Electronics Engineers to convene such collaborations.[41] Although these efforts are critical and should be sustained, government involvement can address market failures that create obstacles to adequate investment in the research and implementation of AI safety and security measures. Greater focus at the federal level will also help bridge gaps between industry and governmental programming.

Over time, discussion on critical issues of testing, evaluation, verification, and validation could extend to security dialogues and future military-to-military engagements. The United States, China, and Russia whether bilaterally or multilaterally in conjunction with other major militaries, could explore the development of policy and technical standards for the robustness and assurance of AI systems used for military purposes, including best practices for testing, evaluation, verification, and validation, pursuant to requirements. For example, security experts and militaries could agree that no highly automated or autonomous systems should be involved in targeting decisions where there is any risk of civilian targets being harmed, unless and until it meets specific legal and robust technical standards. These standards could include establishing with sufficient confidence whether a given system can appropriately distinguish between military and civilian targets in compliance with international humanitarian law. On the one hand, militaries around the world have been integrating semi-automated functionalities into weapons systems since the 1980s, especially for air defense, without robust technical standards in place.[42] On the other hand, existing weapons systems with varying degrees of automation or autonomy have experienced notable failures and accidents in correctly identifying valid targets.[43]

The legal and policy issues that arise from these technical concerns could be addressed by sharing and clarifying policy frameworks. These efforts could also extend to questions of legal liability, policy parameters, and the application of the laws of armed conflict frameworks to the use of AI-enabled or autonomous weapons systems. For instance, the U.S. Department of Defense's Directive 3000.09 on Autonomy in Weapons Systems established

guidelines "designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements," including policies for testing, evaluation, verification, and validation; training; doctrine; and tactics, techniques, and procedures.[44] It is unknown whether the Chinese or Russian militaries have developed or plan to produce comparable policies.

5. Integrate AI safety and security concerns into existing U.S.-China and U.S.-Russia strategic dialogues on cyber security, nuclear issues, and strategic stability.

The United States, China, and Russia should consider incorporating issues of AI safety and security into ongoing dialogues on cybersecurity and nuclear stability.[45] The integration of AI across a range of military applications will impact dynamics in the cyber domain and may even have relevance to nuclear capabilities. The impact of AI on global security will depend on how it is integrated into and employed within existing domains and systems.[46] In particular, there are reasons for concern about the impact of autonomous weapons systems on the future of deterrence and strategic stability.[47] Debate continues about the potential for a "new era" of counterforce to undermine the survivability of nuclear arsenals.[48] The introduction of AI could improve targeting by enabling a degree of reliability, precision, and coordination in detection and targeting not previously possible, rendering nuclear arsenals potentially vulnerable in novel ways that could prove destabilizing.[49]

Against the backdrop of strategic technological uncertainties, various militaries may pursue divergent approaches or behave in accordance with distinct calculations of risk. Militaries tend to evaluate each other's intentions and capabilities in terms of worst-case possibilities. Given this reality, and the likely deficit of trust on both sides, security dialogues can serve a critical function by ensuring regular communication, including the identification and mitigation of biases, risks, and misperceptions. These efforts could also include discussions of approaches to maintaining meaningful human control of AI-enabled and autonomous weapons systems, as well as specific options for crisis management and de-escalation in response to accidents or systemic failures in military applications of AI/ML.[50] It will also be important to recognize and seek clarification on differences in U.S. and Chinese nuclear doctrines, including apparent differences in the weighting ascribed to false positives and false negatives,[51] as well as differing approaches to crises and escalation management.[52]

Ongoing dialogues could contribute to a shared understanding of the technical risks and possibilities of unintended engagements and escalatory consequences with greater autonomy and increased employment of AI/ML techniques, such as in cyber capabilities and operations. Policymakers should weigh the relative benefits of working through existing mechanisms for dialogue as opposed to creating new multilateral approaches and standalone dialogues. In recent history, while concrete progress on risk reduction and crisis management has proven more elusive, U.S.-China military-to-military relations have achieved some success in conveying signals and clarifying intentions.[53] It will be all the more critical to seek and reach shared understandings on these issues in the future.

6. Devise potential parameters and institutional architectures for an "open skies on AI."

In moments of intense rivalry, improvements in transparency can reduce misperceptions and promote shared situational awareness in ways that render clandestine preparations for a military attack more challenging. During the Cold War, the United States proposed to the Soviet Union what later became known as the Open Skies Treaty.[54] While the Soviets initially rebuffed this proposal, it helped lay the groundwork for verification regimes that the United States and Soviet Union agreed to within their nuclear treaties. The Open Skies concept was subsequently revived and negotiated between members of NATO and the Warsaw Pact.[55] Signed in 1992, the treaty allows for aerial surveillance of military forces and activities.[56]

Recently, the future of "Open Skies" has been called into question, in part because of challenges in the treaty's implementation, but defenders underline its historical and continuing importance as a model for promoting transparency through agreed mechanisms for monitoring and information sharing.[57] In AI, verification may prove far more challenging, but technical solutions, including those that enable sharing of the characteristics of a model without compromising the privacy of the training data used, could facilitate an appropriate balancing of security and transparency.[58] As attention turns to AI as a new focal point of rivalry, the uncertainty of measuring relative progress in AI research and its military applications could heighten the security dilemmas that often characterize and exacerbate great power rivalries.[59] Militaries may tend toward overestimating or exaggerating each other's capabilities, even to an extent that can fuel arms racing dynamics, while potentially overlooking the risks of misperception or miscalculation.

For AI safety, one way to manage this dilemma could be for technical and non-technical experts from both countries to conduct shared demonstrations of commercial AI systems and observe the systems' failure modes. These shared demonstrations could build trust, foster a culture of responsibility among AI researchers and organizations, and spur research into AI safety protocols. Since commercial enterprises account for the majority of AI research and spending, including research and spending relevant to future military capabilities, policymakers should increase their outreach to companies, particularly those openly contributing to dual-use or military research and applications. Governments should also dedicate resources to sectoral analysis of how commercial technologies are governed and managed, including to prevent their exploitation by non-state actors with malicious intentions.[60] For instance, it could be productive to develop norms and guidelines on adversarial examples and data poisoning, or at least an improved understanding of vulnerabilities in AI/ML systems.

7. Establish channels to share AI research whose transfer and diffusion could prove generally beneficial.

Under certain circumstances, it may be mutually beneficial to transfer—even to rivals or potential adversaries—technologies or techniques to reduce risk and prevent accidents. During the Cold War, the United States developed and offered to share permissive action links as a cryptographic control to guard against unauthorized employment of nuclear weapons.[61] At present, a comparable undertaking could include efforts to define the types of AI research both countries would be willing to share and promulgate. Experts from the United States, China, and Russia could explore improvements in AI safety and surety, such as failsafe mechanisms or supervisory algorithms. Of course, there is a risk that sharing these ideas could be one-sided or subject to exploitation, but initial exchanges on the topic could gauge the viability of this approach.

AI research is often open and open source in character, but established channels can facilitate the transfer of knowledge in areas of mutual benefit. For example, advances in interpretability could allow researchers to better understand and anticipate risks in the AI systems they are building and deploying. Even among competitors, there may be benefits to sharing and collaborating on machine learning models that provide estimates of uncertainty. If major militaries or intelligence communities become more reliant on AI/ML for early warning and predictive analytics, for instance, potential mistakes could create dynamics of algorithmic misperceptions that result in strategic miscalculation or accidental escalation.[62] There may also be

benefits in exchanging best practices and sharing information on progress in developing countermeasures to adversarial examples; detection of bugs, ambiguities, and negative externalities; and scenario planning for misuses of or accidents involving AI. Such exchanges could happen, and to some degree are already occurring, in industry and academia because of the relative openness of AI as a field. Governments can play a role in facilitating these discussions, connecting them to ongoing policy processes, and establishing parameters to mitigate risk in the process.

In the future, policymakers could explore the development and exchange among militaries of failsafe mechanisms to prevent loss of control or provide a "circuit-breaker" or failsafe in the case of unintended engagements. At a minimum, governments could agree on a protocol for informing each other of AI-enabled or autonomous systems that are malfunctioning, such as an underwater autonomous system that experiences navigation difficulties resulting in its intrusion into another state's territorial waters. One template for such a protocol could be the Incidents at Sea Agreement between the United States and the Soviet Union.[63] It will be important to address the potential for gaps in development and assurance from a multidisciplinary perspective.[64] Already, major power militaries are developing weapons systems with varying degrees of autonomy that incorporate new advances in AI/ML. Given the many reasons for skepticism that a ban is feasible in the near future, the introduction of failsafe and crisis management mechanisms could reduce risks.

Lessons Learned and Challenges

Pragmatic engagement on these core concerns of AI safety, security, and stability must be informed by an understanding of past experiences and potential challenges. The lessons learned from previous U.S.-China and U.S.-Russia dialogues and technological collaborations can maximize the areas of productive conversation and minimize the potential negative consequences of these complex interactions. At the same time, the choice between pursuing bilateral and multilateral engagement on these issues merits careful consideration. As Track 2 dialogues in the Asia-Pacific have expanded significantly in scope and number, the research on and evaluation of their impact remains nascent.[65] For the United States, direct dialogue with China and Russia may be productive on some issues, yet bilateral efforts have at times encountered intense frictions. Multilateral engagements may therefore be beneficial on issues of more general concern. These dynamics deserve continued study.

We encourage consideration of the following lessons learned and principles to inform ongoing and future initiatives:

- *Pursue a practical approach to progress and aim to achieve realistic objectives.* There is value in tackling narrow, concrete issues that can show early, tangible results in order to sustain momentum for continued discussions, particularly at a moment of intensifying geopolitical competition. In pursuing a long-term vision, a persistent focus on initial steps and concrete deliverables is important to manage expectations, since results tend to require patience and "continuous engagement" over time.[66]

- *Convene relevant stakeholders with the right range of expertise, experience, and perspectives.* It is vital to bring the right participants to the table from the start, ensuring a good match of counterparts with relevant expertise and experience on both sides. Ideally, the process of convening the dialogue should be informed by mapping out the key players and stakeholders in AI safety and security on both or all sides, identifying critical needs, misperceptions, and productive avenues for conversation. Seniority can be important to ensure that the outcomes have buy-in from relevant stakeholders. At the same time, ensuring current expertise and relevant experience requires looking beyond the "usual suspects" or typical interlocutors.[67] In particular, dialogues on AI and otherwise should consider diversity and inclusion as core factors that mitigate groupthink in these conversations.[68] From a practical perspective, organizers should also take into account geographic considerations that may affect various participants differently, including due to security concerns and difficulty in obtaining visas or permission to travel. Given security concerns among the United States, China, and Russia, there is a trend toward having dialogues in third countries, elsewhere in Asia or in Europe. This is one practical approach to maximizing participation on both or all sides.

- *Expect and require reciprocity and symmetry in exchange.* In some past dialogues, perceived disparities in levels of openness and transparency among Chinese and American participants have tended to become a source of friction. Indeed, while sharing and signaling on important messages can be productive regardless, limited progress in transparency can undermine the sustainability of engagement, or even raise questions about attempts at deliberate misdirection or manipulation through misinformation.[69] Ideally, these efforts should be

designed in a manner that requires and is contingent on relative reciprocity and promotes symmetry in exchanges.

- *Mitigate the risks of technology transfer and counterintelligence.* U.S.-China dialogues and technical exchanges have often provoked concerns about their potential for exploitation of access to sensitive information and targeting of intelligence efforts against experts, scientists, or officials.[70] It is important to consider and balance the tradeoffs: maximizing the granularity of these conversations by including more strategic and technical experts, on the one hand, and mitigating the risks of collection and counterintelligence on the other. Although the state of AI as a field remains open and collaborative, participants in these dialogues should exercise caution and appropriate judgment to address concerns about dual-purpose exploitation of knowledge or technology. At the same time, a norm or expectation that both/all sides will be providing reporting through appropriate channels to their respective governments can also be productive in ensuring that insights and lessons learned are conveyed.

    - On both sides, participants may be concerned about their counterparts' objectives, potential engagement in collection, or both. In some cases, Chinese participants in such engagements may be linked to and could enable targeting for intelligence purposes, including for the Ministry of State Security or military intelligence. The Chinese Communist Party's United Front Work Department has also leveraged dialogues as a means of cultivating relationships and shaping perceptions with the objective of exerting influence.[71]

    - For American participants, preparation is paramount, and awareness of issues like personal cyber security is critical. For instance, the U.S. government should dedicate more resources and establish clearer parameters for informing and pre-briefing military and government representatives who are participating in these dialogues, as well as civilian, non-governmental participants.

- *Ensure that dialogues and collaborative engagements are structured, routinized, and regularly evaluated for their results.* It is important to ensure that Track 2 conversations are regular and ongoing as opposed to one-off events. This consistency may require developing concrete metrics to evaluate results and justify continuation. For

instance, producing reports or joint statements and analytic or academic publications can be constructive, but these products must be balanced against the value of patience and confidentiality in discussions.

- *Establish clear communication and address the potential for misinterpretation.* Initiatives should address difficulties in miscommunication or misinterpretation issues up front and early, such as clarifying definitions of "AI," AI safety and security, and standards for robustness and assurance, as well as best practices for testing, evaluation, verification, and validation. Practically, reliable simultaneous interpretation, which often requires specialized expertise in translation on topics that can be highly technical, is a prerequisite for productive conversations.

- *Prioritize candor in raising serious concerns, while maintaining civility and amicable interaction.* It is counterproductive to refrain from fully and openly articulating urgent concerns and differences of opinion, including on issues of values and human rights, at a time when serious issues remain unresolved in U.S.-China and U.S.-Russia relations. In particular, U.S.-China engagement can risk legitimizing human rights atrocities in Xinjiang, civil rights violations in Hong Kong, and other deeply concerning practices unless participants raise them in the conversation where salient.[72] The risks of moral hazard and "ethics-washing" are real, particularly if dialogues on issues of AI ethics, safety, and security in China fail to address the ways in which the Chinese government has been leveraging AI to bolster state capacity for censorship and surveillance, including the use of facial recognition to target ethnic minorities.[73] Fear of causing offense or a tendency toward self-censorship can hinder those at the table from being willing to raise such issues directly and without fear of "losing access" or being subject to retaliation (e.g., being denied a visa).[74]

- *Coordinate and maintain shared situational awareness across related dialogues.* Consistently, a lack of adequate coordination across or among dialogues and within the U.S. government can hinder the development of shared situational awareness and promulgation of lessons learned. These dialogues are often not public, but there is a norm and an expectation that the sides will brief their respective governments, which is also important to ensure that results can inform current and future Track 1 interactions. No single clearinghouse in the Department of State, Department of Defense, or elsewhere in the U.S.

government appears to track and monitor these activities. There are often major asymmetries in coordination and preparation among American, Chinese, and international participants in this regard. As a consequence, the U.S. government may have limited visibility on what's happening and where Track 2s have a logical tie-in with Track 1 initiatives. There should be tighter feedback loops between Track 1 and Track 2 dialogues where appropriate, including meetings and coordination among governmental and non-governmental stakeholders throughout the process to ensure clarity of objectives, information sharing, and channels for actionable recommendations.

- *Examine and introduce lessons learned from recent and historical experiences.* There is value in undertaking more comprehensive assessments of different Track 2 conversations and diplomatic negotiations on arms control to identify lessons learned and best practices. It is important to capture and operationalize lessons learned from the recent history of Track 1.5/2 dialogues with China and Russia, particularly those involving nuclear issues and, more recently, cyber security, as well as ongoing engagements with allies and partners.[75]

Conclusions and Implications

The stakes are too high to refrain from pursuing challenging conversations on AI safety and security. It is encouraging that policymakers and stakeholders in the United States, China, Europe, and worldwide appear to be open to and support engagement in these discussions.

Against the backdrop of complex geopolitical contingencies, trust among great powers is lacking, and even historical alliances are coming under new pressure. These issues transcend any single country or government, involving a much wider range of stakeholders and uniquely complex threats and challenges. Even if progress is slow, the benefits include potential for instilling common norms, practices, and behaviors within discrete communities.[76]

On these vital issues, pragmatic engagement should pursue courses of action that can be productive even in the case of a deficit or absence of trust. U.S.-China competition has been heightened by the mutual construction of narratives that reinforce rivalry and adverse perceptions of "the Other."[77] Attempts to build trust or address mistrust are unlikely to yield near-term dividends, given the overall climate of great power relations. In this context,

policymakers will need to explore options for costly signaling of intention and credibility.[78] It is worth considering the range of possible signals that both sides might convey to enhance the credibility of their respective commitments on AI safety and security. All militaries should consider restraint in deployment of AI-enabled and autonomous weapons systems while exploring mechanisms to demonstrate commitment and implementation of any agreements in the absence of feasible options for verification.

The shared threats and challenges of AI present an opportunity for pragmatic international engagement. In recent history, the United States has been a leader in the construction and maintenance of the rules-based international order. New trends in geopolitics and emerging technologies present new challenges to existing institutions and demand new paradigms to mitigate risk. As today's order comes under greater pressure, it is imperative for American policymakers to re-engage on such critical challenges. U.S. policy priorities on AI safety begin at home, but also must extend to international engagement and collaboration with allies, partners, and competitors. The core issues of AI security are at once abstrusely technical and incredibly strategic, demanding attention at the highest levels of leadership. Creative thinking on policy solutions can look to historical lessons from bilateral and multilateral engagements on military-technical issues, while seeking novel solutions. Great powers should exercise greater agency in shaping the future of AI and responding to the dilemmas it poses for global security and stability.

Acknowledgments

The analysis and recommendations in this paper are based on the authors' experiences participating in a number of Track 1.5 and Track 2 dialogues and exchanges that have addressed technology issues, cyber conflict, artificial intelligence, and general concerns in U.S.-China relations. Elsa Kania would like to thank and acknowledge the Chinese, American, and international colleagues with whom she had the chance to interact and learn from in the course of these engagements.

Both authors are grateful to two external reviewers, Eric Richardson and Danit Gal, for their insightful comments and perspective. Zachary Arnold, Ryan Fedasiuk, Carrick Flynn, Roxanne Heston, Saif Khan, Margarita Konaev, Jason Matheny, Michael Page, Helen Toner, Igor Mikolic-Torreira, Alexandra Vreeman, and Lynne Weil provided helpful feedback on this paper at various stages. All errors are the authors' own.

Document Identifier: 10.51593/20190051

Endnotes

[1] U.S. and Chinese AI plans, policies, and leadership statements have consistently emphasized this point. See, e.g., State Council Notice on the Issuance of the New Generation AI Development Plan" [国务院关于印发新一代人工智能发展规划的通知], July 20, 2017, http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm; Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania, "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)," New America, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/; White House, "China's New Executive Order on Maintaining American Leadership in Artificial Intelligence," February 11, 2019, https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/; "Xi Jinping: Promote the Healthy Development of Our Nation's New Generation of Artificial Intelligence" [习近平：推动我国新一代人工智能健康发展], *Xinhua*, October 31, 2018, http://www.xinhuanet.com/politics/2018-10/31/c_1123643321.htm. For an English translation, see Elsa Kania and Rogier Creemers, "Xi Jinping Calls for 'Healthy Development' of AI (Translation)," New America, November 5, 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/xi-jinping-calls-for-healthy-development-of-ai-translation/.

[2] Klaus Schwab, *The Fourth Industrial Revolution* (Geneva: World Economic Forum, 2016).

[3] Some scholars and policymakers, including former U.S. Secretary of Defense James Mattis, have posited that AI could alter not only the character but even the nature of warfare, while other observers remain more skeptical that AI will prove that transformative. See "Artificial intelligence poses questions for nature of war: Mattis," *Phys.org*, February 18, 2018, https://phys.org/news/2018-02-artificial-intelligence-poses-nature-war.html. For an academic perspective on the topic, see Frank G. Hoffman, "Will War's Nature Change in the Seventh Military Revolution?" *Parameters* 47, no. 4 (2017): 19-31.

[4] Michael C. Horowitz, Gregory Allen, Elsa Kania, and Paul Scharre, "Strategic Competition in an Era of Artificial Intelligence," Center for New American Security, August 2018.

[5] Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, May 2018.

[6] See, e.g., "National Security Strategy of the United States of America," December 2017, https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf.

[7] See, e.g., Todd S. Sechser, Neil Narang, and Caitlin Talmadge, "Emerging technologies and strategic stability in peacetime, crisis, and war," *Journal of Strategic Studies* 42, no. 6 (2019): 727-735. There has been a range of articles written either describing or debunking the notion that there is an "AI arms race" underway. See, e.g., Edward Moore Geist, "It's already too late to stop the AI arms race—We must manage it instead," *Bulletin of the Atomic Scientists* 72, no. 5 (2016): 318-321; Elsa B. Kania, "The Pursuit of AI Is More Than an Arms Race," *Defense One*, April 18, 2018, https://www.defenseone.com/ideas/2018/04/pursuit-ai-more-arms-race/147579/; Remco Zwetsloot, Helen Toner, and Jeffrey Ding, "Beyond the AI Arms Race: America,

China, and the Dangers of Zero-Sum Thinking," *Foreign Affairs*, November 16, 2018, https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race; Heather Roff, "The frame problem: The AI 'arms race' isn't one," *Bulletin of the Atomic Scientists*, April 29, 2019, https://thebulletin.org/2019/04/the-frame-problem-the-ai-arms-race-isnt-one/.

[8] Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, May 2018.

[9] "Drones, robots, lasers, supersonic gliders & other high-tech arms: Putin wants Russian military to be up to any future challenge," *Russia Today*, November 22, 2019, https://www.rt.com/russia/474119-putin-laser-drone-robot-hypersonic/.

[10] Amanda Macais, "Russia's nuclear-armed underwater drone may be ready for war in eight years," *CNBC*, March 25, 2019, https://www.cnbc.com/2019/03/25/russias-nuclear-armed-underwater-drone-may-be-ready-for-war-in-2027.html.

[11] Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs*, 98 (2019): 135.

[12] John Hawley, "Patriot Wars," Center for a New American Security, January 2017.

[13] See, e.g., Huang Ling, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. Doug Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and Artificial Intelligence*, ACM (2011): 43-58; Wieland Brendel, Jonas Rauber, and Matthias Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248* (2017); Dong Yinpeng, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018): 9185-9193.

[14] Audrey K. Cronin, *Power to the People: How Open Technological Innovation is Arming Tomorrow's Terrorists* (Oxford: Oxford University Press, 2019).

[15] For one example of Chinese military writings on the topic available in English, see Jian Zhou, *Fundamentals of Military Law: A Chinese Perspective* (New York: Springer, 2019).

[16] See "Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts" (Protocol I), June 8, 1977, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/WebART/470-750045. For an analysis of the challenges of applying this process to weapons systems with increasing autonomy, see Vincent Boulanin, *Implementing Article 36 Weapon Reviews in the Light of Increasing Autonomy in Weapon Systems* (Stockholm International Peace Research Institute [Stockholms internationella fredsforskningsinstitut]: SIPRI, 2015).

[17] For one relevant framing and scoping of the topic, see Pedro Ortega and Vishal Maini, "Building safe artificial intelligence: specification, robustness, and assurance," *DeepMind Safety Research Blog*, 2018).

[18] By some estimates, less than one percent of AI research and development funding is directed to AI security at present. See "The 3 major security threats to AI," *c4isrnet*, https://www.c4isrnet.com/artificial-intelligence/2019/09/10/the-3-major-security-threats-to-ai/. However, it is worth noting that the field is rapidly expanding in ways that are promising, with support and engagement from stakeholders that include OpenAI, the Future of Life Institute, and the Future of Humanity Institute, among others.

[19] For instance, there are specific research programs that are dedicated to the technical aspects of these issues, but lesser progress seemingly on policy frameworks. See, e.g., "Trojans in Artificial Intelligence (TrojAI)," https://www.iarpa.gov/index.php?option=com_content&view=article&id=1150&Itemid=448.

[20] The authors are indebted to Rita Konaev and Sam Bendett for their insights on these issues. Margarita Konaev and Samuel Bendett, "Russian AI-Enabled Combat: Coming to a City Near You?", *War on the Rocks*, July 31, 2019, https://warontherocks.com/2019/07/russian-ai-enabled-combat-coming-to-a-city-near-you.

[21] China Institute of Information and Communications Technology (CAICT), "AI Security White Paper" [人工智能安全白皮书], September 2018, http://www.caict.ac.cn/kxyj/qwfb/bps/201809/P020180918473525332978.pdf. See also this English translation and evaluation of the white paper: Elsa Kania, Dahlia Peterson, Lorand Laskai, and Graham Webster, "Translation: Key Chinese Think Tank's "AI Security White Paper" (Excerpts)," *DigiChina*, February 21, 2019 https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-key-chinese-think-tanks-ai-security-white-paper-excerpts/.

[22] Ibid.

[23] For the English translation of the Beijing AI Principles, see "Beijing AI Principles," May 28, 2019, https://www.baai.ac.cn/blog/beijing-ai-principles. " "'Beijing Consensus on Artificial Intelligence' released 15 principles to regulate AI research and development, employment and governance" [《人工智能北京共识》发布 15条原则规范AI研发、使用和治理], China News Network [中国新闻网], May 25, 2019, http://www.xinhuanet.com/local/2019-05/25/c_1124541344.htm.

[24] This term, which could also be translated as "social occupations," appears to imply concern with the social and societal impact of AI, including on patterns of employment and the relevance of various occupations, such as whether AI could result in mass unemployment.

[25] "China's first AI security and rule of law guidelines released" [国内首个人工智能安全与法治导则发布], Guangming Network [光明网], August 30, 2019, http://tech.gmw.cn/2019-08/30/content_33123802.htm.

[26] On the notion of "robust AI," see, e.g., the research that is occurring under the aegis of DARPA, "AI Next Campaign," https://www.darpa.mil/work-with-us/ai-next-campaign.

[27] For relevant antecedents, see the past U.S.-China initiative to develop a glossary of terminology on nuclear security that was created through engagement between the National

Academies of Science Committee on International Security and Arms Control and its Chinese counterparts, which was published in Washington, D.C. and Beijing. See *English-Chinese Chinese-English Nuclear Security Glossary* (Washington, D.C.: National Academies Press, and Beijing: Atomic Energy Press, 2008).

[28] For instance, there may be feasible options for collaboration on shared threats at the intersection of AI and cyber security, such as the potential intensification of cyber-criminal activities.

[29] Jennifer Bouey, "Implications of US-China Collaborations on Global Health Issues," Testimony presented before the U.S.-China Economic and Security Review Commission," July 31, 2019, https://www.uscc.gov/sites/default/files/Bouey%20Written%20Statement.pdf.

[30] For an excellent and impactful example, see Shu Yuelong, Ying Song, Dayan Wang, Carolyn M. Greene, Ann Moen, C. K. Lee, Yongkun Chen et al, "A ten-year China-US laboratory collaboration: improving response to influenza threats in China and the world, 2004–2014," *BMC Public Health* 19, no. 3 (2019): 520.

[31] Chris Cox, "The Cox Report: US National Security and Military/Commercial Concerns with the People's Republic of China," Vol. 105, no. 851, Regnery Publishing, 1999.

[32] Aaron Mehta, "How the US and China collaborated to get nuclear material out of Nigeria — and away from terrorist groups," *Defense News*, January 14, 2019, https://www.defensenews.com/news/pentagon-congress/2019/01/14/how-the-us-and-china-collaborated-to-get-nuclear-material-out-of-nigeria-and-away-from-terrorist-groups/.

[33] For context, see "Apollo-Soyuz Test Project Overview," NASA, https://www.nasa.gov/apollo-soyuz/overview.

[34] There is fascinating literature on these issues, including Anna Geltzer, "In a distorted mirror: the Cold War and US-Soviet biomedical cooperation and (mis) understanding, 1956–1977," *Journal of Cold War Studies* 14, no. 3 (2012): 39-63; Glenn E. Schweitzer, *Techno-Diplomacy: US-Soviet Confrontations in Science and Technology* (New York: Plenum Press 1989).

[35] See, e.g., Ryan Hass and Zach Balin, "US-China relations in the age of artificial intelligence," Brookings Institution, January 10, 2019, https://www.brookings.edu/research/us-china-relations-in-the-age-of-artificial-intelligence/.

[36] For one example of recent topics of joint research, see "Dear Colleague Letter: NSF/NSFC Joint Research on Environmental Sustainability Challenges," National Science Foundation, July 20, 2018, https://www.nsf.gov/pubs/2018/nsf18096/nsf18096.jsp.

[37] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565* (2016).

[38] "Pugwash Conferences on Science and World Affairs: Facts," The Nobel Prize, 1995, https://www.nobelprize.org/prizes/peace/1995/pugwash/facts/.

[39] For context, see Mike Moore, "Forty years of Pugwash," *Bulletin of the Atomic Scientists* 53, no. 6 (1997): 40-45.

[40] For context on Track 1.5 approaches, see Oliver Wolleh, "Track 1.5 Approaches to Conflict Management: Assessing Good Practice and Areas for Improvement," Berghof Foundation for Peace Support, 2007, https://peacemaker.un.org/sites/peacemaker.un.org/files/Track1.5ApproachestoConflictManagement_BerghofFoundation2007.pdf.

[41] See, e.g., "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

[42] Paul Scharre and Michael C. Horowitz, "An Introduction to Autonomy in Weapons Systems," Center for a New American Security, February 2015, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.

[43] For salient examples, see John Hawley, "Patriot Wars," Center for a New American Security, January 2017; Gene I. Rochlin, "Iran Air Flight 655 and the USS Vincennes," in *Social Responses to Large Technical Systems* (Springer, Dordrecht, 1991), 99-125.

[44] See "Directive 3000.09 on Autonomy in Weapons Systems," https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[45] See, e.g., Michael O. Wheeler, "Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned," No. IDA-P-5135, *Institute for Defense Analyses*, 2014, https://apps.dtic.mil/dtic/tr/fulltext/u2/a622481.pdf; "Track 1.5 U.S.-China Cyber Security Dialogue," https://www.csis.org/programs/technology-policy-program/cybersecurity-and-governance/other-projects-cybersecurity/track-1; Eli Lake, "Managing Cyberwar With Vodka," *Bloomberg*, January 4, 2019, https://www.bloomberg.com/opinion/articles/2019-01-04/managing-cyberwar-with-vodka-russia-china-u-s-meet.

[46] Such dialogues could explore ways in which AI contributes to collaborative initiatives on cyber security and nuclear security.

[47] Michael C. Horowitz, "When Speed Kills: Autonomous Weapon Systems, Deterrence, and Stability," *Journal of Strategic Studies* 42, Issue 6 (2019): 764-788.

[48] Keir A. Lieber and Daryl G. Press, "The new era of counterforce: Technological change and the future of nuclear deterrence," *International Security* 41, no. 4 (2017): 9-49.

[49] For a notable analysis of this issue, see, for instance: Rafael Loss, "Artificial Intelligence, the Final Piece to the Counterforce Puzzle?" No. LLNL-TR-791947 (Lawrence Livermore National Lab, Livermore, CA: United States, 2019).

[50] For a more detailed discussion of these concepts, see Heather M. Roff, "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous

Weapons," in Briefing paper prepared for the Review Conference of the Convention on Conventional Weapons, 2016.

[51] See, e.g., Lora Saalman, "Fear of False Negatives: AI and China's Nuclear Posture," *Bulletin of the Atomic Scientists*, April 24, 2018, https://thebulletin.org/2018/04/fear-of-false-negatives-ai-and-chinas-nuclear-posture/. For a more detailed assessment, see Lora Saalman, "China's Integration of Neural Networks into Hypersonic Glide Vehicles," December 2018, https://nsiteam.com/social/wp-content/uploads/2018/12/AI-China-Russia-Global-WP_FINAL.pdf.

[52] Fiona S. Cunningham and M. Taylor Fravel, "Dangerous Confidence? Chinese Views on Nuclear Escalation," *International Security* 44, no. 2 (2019): 61-109.

[53] Scott W. Harold, "Optimizing the US-China Military-to-Military Relationship," *Asia Policy* 26, no. 3 (2019): 145-168.

[54] The concept was proposed in 1955 by President Eisenhower, but it was rejected by the Soviet Union at the time.

[55] For context, see Daryl Kimball, "The Open Skies Treaty at a Glance," Arms Control Association, October 2019, https://www.armscontrol.org/factsheets/openskies.

[56] For context, see James J. Marquardt, "Transparency and Security Competition: Open Skies and America's Cold War Statecraft, 1948–1960," *Journal of Cold War Studies* 9, no. 1 (2007): 55-87.

[57] Leonid Bershidsky, "U.S. and Russia Should Keep the Skies Open," *Bloomberg*, October 30, 2019, https://www.bloomberg.com/opinion/articles/2019-10-30/u-s-and-russia-must-keep-the-open-skies-treaty; Alexandra Bell and Anthony Wier, "Open Skies Treaty: A Quiet Legacy Under Threat," *Arms Control Today*, January/February 2019, https://www.armscontrol.org/act/2019-01/features/open-skies-treaty-quiet-legacy-under-threat.

[58] For relevant technical literature, see Mohassel, Payman, and Yupeng Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE (2017): 19-38.

[59] Adam Breuer and Alastair I. Johnston, "Memes, narratives and the emergent US–China security dilemma," *Cambridge Review of International Affairs* 32, no. 4 (2019): 429-455.

[60] See, e.g., Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe et al. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228* (2018).

[61] For context, see Dan Caldwell, "Permissive action links: a description and proposal," *Survival* 29, no. 3 (1987): 224-238; Peter Stein and Peter Feaver. *Assuring control of nuclear weapons: The evolution of permissive action links*. No. 2. (University Press of America, 1987). See also Steven M. Bellovin, "Permissive Action Links, Nuclear Weapons, and the History of Public Key Cryptography," https://web.stanford.edu/class/ee380/Abstracts/060315-slides-bellovin.pdf.

[62] For instance, if an algorithm used in early warning indicates mistakenly that an attack is imminent or underway, that erroneous conclusion could prove destabilizing.

[63] David F. Winkler, "The Evolution and Significance of the 1972 Incidents at Sea Agreement," *Journal of Strategic Studies* 28, no. 2 (2005): 361-377.

[64] For a good example of this emerging literature, see Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter, "Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective," *Artificial Intelligence* (2019): 103-201.

[65] Desmond Ball, Anthony Milner, and Brendan Taylor, "Track 2 Security Dialogue in the Asia-Pacific: reflections and future directions," *Asian Security* 2, no. 3 (2006): 174-188.
[66] Michael O. Wheeler, "Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned," No. IDA-P-5135, Institute for Defense Analyses, Alexandria, VA, 2014, https://apps.dtic.mil/dtic/tr/fulltext/u2/a622481.pdf.

[67] Not only technologists but also social scientists have much to contribute to these conversations, considering the critical human element. See "AI Safety Needs Social Scientists," Open AI, February 19, 2019, https://openai.com/blog/ai-safety-needs-social-scientists/.

[68] In negotiations and conflict resolution, diversity and inclusion of all kinds are important. For instance, research on peace processes shows that when women are engaged as mediators and negotiators, it is more likely that peace agreements will endure. See Jana Krause, Werner Krause, and Piia Bränfors, "Women's Participation in peace negotiations and the durability of peace," *International interactions* 44, no. 6 (2018): 985-1016. On the participation and engagement of stakeholders in civil society, see Zanker, Franzisca. "Legitimate representation: Civil society actors in peace negotiations revisited," *International Negotiation* 19, no. 1 (2014): 62-88.

[69] For instance, in some dialogues, American participants tend to be far more forthcoming about their strategy, doctrine, and technological developments. It may be beneficial for the United States to clarify potential Chinese misconceptions of persistent engagement, but less productive when China refuses to acknowledge even basic, established information about its own cyber forces and doctrinal concepts.

[70] These concerns are not new, but rather date back to the Cox Report. See "The Cox Report: US National Security and Military/Commercial Concerns with the People's Republic of China," Vol. 105, no. 851. Regnery Publishing, 1999.

[71] For context on China's united front work, see Alexander Bowe, "China's Overseas United Front Work: Background and Implications for the United States," US-China Economic and Security Review Commission, 2018. For an early perspective on the topic, see Larry M. Wortzel, "Why Caution Is Needed In Military Contacts with China," Heritage Foundation, 1999.

[72] See, e.g., the following Human Rights Watch report on China and the thoroughly sourced and expanding literature on crimes against humanity in Xinjiang: "China: Events of 2018," Human Rights Watch, https://www.hrw.org/world-report/2019/country-chapters/china-

and-tibet. See also the myriad academic and analytical writings that attest to the situation: Adrian Zenz, "Beyond the Camps: Beijing's Grand Scheme of Coercive Labor, Poverty Alleviation and Social Control in Xinjiang," Testimony to the U.S.-China Economic and Security Review Commission, 2019; Adrian Zenz, "Break Their Roots: Evidence for China's Parent-Child Separation Campaign in Xinjiang," *Journal of Political Risk* 7, no. 7 (2019); James Leibold, "Surveillance in China's Xinjiang Region: Ethnic Sorting, Coercion, and Inducement," *Journal of Contemporary China* (2019): 1-15.

[73] For authoritative reporting on these issues, see Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *New York Times*, April 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

[74] For a thoughtful assessment of issues of academic self-censorship, see Sheena C. Greitens and Rory Truex, "Repressive Experiences among China Scholars: New Evidence from Survey Data," *The China Quarterly* (2018): 1-27.

[75] Michael O. Wheeler, "Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned," No. IDA-P-5135. Institute for Defense Analyses, Alexandria, VA, 2014, https://apps.dtic.mil/dtic/tr/fulltext/u2/a622481.pdf.

[76] For an evaluation of these issues in historical perspective, see Alastair I. Johnston, *Social states: China in international institutions*, 1980-2000. Vol. 144 (Princeton: Princeton University Press, 2014).

[77] Adam Breuer and Alastair I. Johnston, "Memes, narratives and the emergent US–China security dilemma," *Cambridge Review of International Affairs* 32, no. 4 (2019): 429-455.

[78] For context and relevant literature, see Andrew H. Kydd, *Trust and mistrust in international relations* (Princeton: Princeton University Press, 2007). Signals are costly when the initiator incurs some risk or expense for making the gesture, thereby revealing information about resolve. This literature on the sending and perception of signals is dynamic and evolving. See, e.g., Kai Quek, "Are costly signals more credible? evidence of sender-receiver gaps," *The Journal of Politics* 78, no. 3 (2016): 925-940.