September 30, 2020

To: National Security Commission on Artificial Intelligence Commissioners and Staff

Comments on the National Security Implications of Artificial Intelligence and Associated Technologies

I. Introduction

The Global Catastrophic Risk Institute (GCRI), the Future of Life Institute (FLI), and the Center for Human-Compatible Artificial Intelligence (CHAI) appreciate the opportunity to inform the final report of the National Security Commission on Artificial Intelligence (the Commission) through our submission of a written comment in response to the Commission's May 2020 solicitation (Docket No. 05-2020-01). Our organizations have collaborated on this response in order to leverage diverse expertise and to highlight the consensus supporting our remarks.

- The Global Catastrophic Risk Institute is a nonprofit and nonpartisan think tank, whose mission is to develop the best ways to confront humanity's gravest threats.
- The Future of Life Institute is a non-profit organization, whose mission is to catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course in considering new technologies and challenges.
- The Center for Human-Compatible Artificial Intelligence is a technical research organization based at the University of California, Berkeley, whose mission is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research toward provably beneficial systems.

Through our work, we have come to share the Commission's view of AI as a set of technologies with important and wide-ranging, perhaps transformative, national security implications. We believe that securing America's people, institutions, and values from the unprecedented challenges associated with these powerful technologies requires broad and sometimes unorthodox thinking that goes well beyond traditional strategic concerns. To that end, and building on the initial and interim reports of the Commission, as well as its white papers and memos, we offer the observations below. We expound on them and recommend concrete policy actions in subsequent sections of this document.

- Managing domestic and international risks associated with artificial intelligence requires an expansive view of national security that accounts for traditional *and* non-traditional concerns.

- Maximizing security requires balancing pursuit of military and technological advantages with recognition of the dangers of international arms races, unintentional conflict escalation, weapons proliferation, harmful accidents, and nonstate malicious use of emerging technologies.
- Maintaining robust diplomatic engagement with allied and non-allied nations alike is necessary to ensure that artificial intelligence enhances rather than undermines security.
- Formulating and implementing effective strategies around AI requires acknowledging the limits of foresight and designing institutions capable of adapting to unanticipated technological and geopolitical developments.
- Ensuring that artificially intelligent systems are safe, reliable, and ethical requires agencies that develop and implement such systems to be transparent and accountable.

For additional information on our observations and recommendations, please contact Jared Brown, Special Advisor for Government Affairs for GCRI, at jared@gcrinstitute.org.

II. Observations and Recommendations

A. **Managing domestic and international risks associated with artificial intelligence requires an expansive view of national security that accounts for traditional *and* non-traditional security concerns.**

Great power competition cannot be the sole focus of U.S. national security thinking in relation to AI. The implications of AI are as wide-ranging—and uncertain—as the technologies themselves. Some AI-related developments promise to fall well within the traditional national security ambit of “defending the homeland, deterring war, protecting allies, and winning on the battlefield”: enhanced geospatial image analysis, improved anti-jamming capabilities, and more rapid decryption, to name a few.[1] Others, however, are poised to affect areas more peripheral to traditional national security paradigms. Example areas include: vaccine research and development, global climate modeling, global availability of high-quality education, and economic development.

As COVID-19 has shown, this second category of developments is at least as integral as the first to the security of America’s people, values, and institutions. To its credit, the Commission has already launched several special projects to address the implications of AI for managing the pandemic and publicized the projects’ findings in its White Paper Series on Pandemic Response and Preparedness. We applaud these efforts. We urge the Commission to go further, however, by incorporating pandemic preparedness and response into its primary body of work and considering in some detail how AI can help address infectious disease and other non-traditional concerns in its final report. Though many other non-traditional national security concerns exist,

[1] National Security Commission on Artificial Intelligence, “Interim Report,” November 2019, p. 29, available at https://drive.google.com/a/nscai.org/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view?usp=sharing.

we highlight the following issues in particular:

- **Infectious disease.** While the Commission has released several papers on infectious disease since the outbreak of COVID-19, preparedness and response to epidemics and pandemics were absent from the Commission's earlier Initial and Interim Reports. With more than 200,000 American deaths and trillions of dollars of economic losses attributable to the disease as of September, these threats continue to remain outside the Commission's primary lines of effort.[2] This is despite the inclusion of biothreats in the 2017 U.S. National Security Strategy and AI's numerous promising applications to pandemic preparedness and response, including "high-performance computing for simulations and other analyses, in support of the design of therapeutics and vaccines, and computational modeling for tracking contagious diseases, monitoring the spread among individuals, predicting future outbreaks, and allocating healthcare resources."[3] Although the COVID-19 pandemic will subside over time as a national security threat, future biological threats are inevitable and could pose greater challenges still for the United States. As the executive director of the Commission writes, "[i]t takes no leap of imagination to envision a similar, or even more catastrophic biothreat emerging from a pathogen engineered for lethality and deployed as a weapon."[4]
- **Climate change.** Rising global temperatures contribute to both immediate and longer-term risks to U.S. national security. According to the U.S. National Oceanic and Atmospheric Administration (NOAA), average temperatures in 2019 were 1.71°F above the twentieth-century average, making it the second warmest year on record. The five warmest years on record have all occurred since 2015.[5] Such trends, which are poised to continue into the foreseeable future, pose serious operational challenges for the United States. In a January 2019 report, for example, the U.S. Department of Defense found that among 79 installations examined, 60 face the prospect of recurrent flooding, 48 face possible drought, and 43 face potential wildfires over the next twenty

[2] For death toll, see: Centers for Disease Control and Prevention, "Excess Deaths Associated with COVID-19," National Center for Health Statistics, https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm. For information about the Commission's lines of effort, see: Eric Horvitz et al., "Privacy and Ethics Recommendations for Computing Applications Developed to Mitigate COVID-19," White Paper Series on Pandemic Response and Preparedness No. 1, National Security Commission on Artificial Intelligence, May 6, 2020, p. 3, available at https://drive.google.com/file/d/1m0AT21dS2XJ6JIGMgo7SuLSLveWIO8WK/view?usp=sharing.

[3] White House, "National Security Strategy of the United States of America," December 18, 2017, p. 9, https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf; Horvitz et al., "Privacy and Ethics Recommendations for Computing Applications Developed to Mitigate COVID-19," p. 4.

[4] National Security Commission on Artificial Intelligence, "The Role of AI Technology in Pandemic Response and Preparedness: Recommended Investments and Initiatives," White Paper Series on Pandemic Response and Preparedness, No. 3, June 25, 2020, p. 3, available at https://drive.google.com/file/d/153DUHToD4zoM_GXe9MWGNKzend7TsI2o/view.

[5] NOAA National Centers for Environmental Information, "State of the Climate: Global Climate Report for Annual 2019," January 2020, https://www.ncdc.noaa.gov/sotc/global/201913.

years—with significant increases in each category.[6] Climate change is also a "threat multiplier" in that it exacerbates geopolitical issues such as large-scale human migration, sociopolitical instability, and supply chain vulnerability. Artificial intelligence offers important opportunities to better understand pertinent dynamics and further mitigation, adaptation, and resilience initiatives by increasing energy efficiency, improving the scope and accuracy of integrated physical/biological/societal models, monitoring agricultural production, improving our ability to evaluate possible interventions, and optimizing city-planning and design, among other applications.[7] Climate-related applications of AI, then, merit significant attention from governments at the federal, state, and municipal levels from both an economic and security standpoint. Furthermore, the United States has a clear interest in reducing greenhouse gas emissions associated with the training and use of AI systems, which can require large amounts of computational resources.[8]

- **Complex systems failures.** The integration of AI into increasing numbers of complex technical systems poses novel safety challenges and opportunities. In certain cases, such as with nuclear command, control, and communications infrastructures; power stations and electrical grids; and the "internet of things," AI-related failures—whether due to malicious intent or accident—can endanger not just human safety but also U.S. national security.[9] Such failures may result from internal malfunctions, unanticipated interactions with surrounding environments and external systems (including adversarial interactions), or misalignment between AI objective functions and human goals. In any case, integrating AI into complex systems requires extreme care. Initial applications of AI in such systems should be engineered with a view toward risk mitigation, including by performing such functions as monitoring for anomalous behavior and alerting personnel of potential dangers.
- **Autonomous weapons systems.** Development and battlefield deployment of semi- and fully autonomous weapons systems (AWS) could entail not only ethical and legal challenges, as are under contention in international fora such as the Convention on Certain Conventional Weapons, but also serious risks to U.S. national security. These include, but are not limited to, the possibility that large anti-personnel drone swarms will be treated or perceived as weapons of mass destruction,[10] dangerous AWS proliferating

[6] Department of Defense, "Report on Effects of a Changing Climate to the Department of Defense," January 2019, p. 5, https://media.defense.gov/2019/Jan/29/2002084200/-1/-1/1/CLIMATE-CHANGE-REPORT-2019.PDF.

[7] World Economic Forum, PwC, and Stanford Woods Institute for the Environment, "Harnessing Artificial Intelligence for the Earth," January 2018, p. 10, http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf

[8] Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650, https://www.aclweb.org/anthology/P19-1355.pdf.

[9] For an example of how complex systems failures can endanger national security, see former Soviet President Mikhail Gorbachev's attribution of the collapse of the Soviet Union to the nuclear accident at Chernobyl: Mikhail Gorbachev, "Turning Point at Chernobyl," *Project Syndicate*, April 14, 2006, https://www.project-syndicate.org/commentary/turning-point-at-chernobyl.

[10] Zachary Kallenborn, "Are Drone Swarms Weapons of Mass Destruction?," Future Warfare Series No. 60, United States Air Force Center for Strategic Deterrence Studies, May 6, 2020.

to non-state actors and becoming a threat to U.S. interests,[11] and AWS provoking rapid escalations in conflicts, potentially leading to "flash war."[12] These and other potential implications require the United States to develop comprehensive strategic guidance on how to prepare for and respond to the national security challenges associated with AWS, particularly in the diplomatic arena of arms control negotiations.

- **Sociopolitical instability.** Unrest, at home and abroad, poses serious challenges for U.S. national security. Artificial intelligence holds the potential to exacerbate or ameliorate these challenges, depending on how the United States and other actors employ—and, in many cases, counteract—pertinent technologies. Malicious actors have already attempted to use AI to influence the outcome of elections and referenda, and to spread disinformation and sow discord.[13] Some analysts contend also that automation involving AI technologies is reconfiguring labor markets and widening inequalities, in turn altering political outcomes.[14] Conversely, authoritarian governments, most notably in China, are using AI to stifle political dissent and exert control over domestic populations, undermining human rights and strengthening anti-democratic regimes whose practices are oftentimes antithetical to American values. Given the potential for sociopolitical trends to factor into geopolitical dynamics, the U.S. national security community would be remiss to neglect AI's applications and implications in relation to these issues.

To address these concerns, we recommend the following actions:

- **Establish positions within the National Security Council (NSC) tasked with interagency coordination of AI-related initiatives.** Establishing these positions, which should include at least one senior director role, within the NSC will promote high-level strategic thinking and reduce stove-piping, increasing the likelihood that both traditional and non-traditional cross-cutting security issues receive due attention. Just as the Joint Artificial Intelligence Center serves as a focal point for AI-related activities in the Department of Defense, dedicated NSC staff would be able to coordinate AI-related activities across the U.S. national security enterprise as a whole. Such staff could, for example, conduct comprehensive cost-benefit analyses of the Commission's recommendations on import/export controls or immigration reforms, which cut across the purview of many agencies.

[11] For example, see: Philip Chertoff, "Perils of Lethal Autonomous Weapons Systems Proliferations : Preventing Non-State Acquisition," Geneva Center for Security Policy, Strategic Security Analysis Paper, Issue 3, May 2018, available at https://www.gcsp.ch/publications/perils-lethal-autonomous-weapons-systems-proliferation-preventing-non-state.

[12] For example, see: Paul Scharre, "A Million Mistakes a Second," *Foreign Policy,* Fall 2018, https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

[13] P.W. Singer and Emerson T. Brooking, *Likewar: The Weaponization of Social Media* (New York: Houghton Mifflin Harcourt, 2018), p. 176.

[14] See, for example: Carl Benedikt Frey, Thor Berger, and Chinchih Chen, "Political Machinery: Did Robots Swing the 2016 U.S. Presidential Election?," *Oxford Review of Economic Policy* 34.3 (July 2018): pp. 418-442, https://academic.oup.com/oxrep/article-abstract/34/3/418/5047377.

- **Issue a new Executive Order on autonomous weapons systems.** Established U.S. policy on AWS is limited in scope to Department of Defense guidance on whether the United States will develop and use AWS,[15] and positions in multilateral fora regarding the applicability of existing international law to AWS.[16] The current paucity of official policies on AWS leaves the United States without a clear strategy for addressing a set of technologies with critical implications for U.S. national security. To remedy this situation, the President of the United States should issue a new executive order containing high-level, comprehensive guidance on issues pertaining to AWS. This guidance should address how the United States will mitigate the risks posed by AWS for arms proliferation, nuclear destabilization, and conflict escalation, among other topics. The E.O. should also order the Department of State to assess the geopolitical implications of potential international regulations on AWS, including their possible prohibition, and clarify national positions on international legal aspects of the technologies. We recognize as well that the Commission seeks "to understand different perspectives on [lethal autonomous weapon systems] and hear from different sides of the issue before attempting to reach consensus judgments."[17] However, in the event that the Commission is unable to reach consensus, the Commission should issue non-consensus findings and recommendations to help inform policy in this critical issue area.
- **Increase funding for applied AI research in the areas of pandemic prevention and response *and* climate change mitigation and adaptation.** The Commission's third white paper on pandemic preparedness and response provides an excellent starting point in this regard, with its recommended investments and initiatives. In particular, we endorse its fourth recommendation: that the United States work to "enhance global cooperation on smart disease surveillance and international health data norms and standards."[18] The Commission should go beyond infectious disease, however, and examine how investments in AI can help address climate change and other ecological challenges, such as biodiversity loss and environmental degradation. As noted in the white paper, climate models have been instrumental to the scientific understanding of disease spread.[19] Using AI to enhance these models, would leave the United States better prepared to handle both environmental and biological threats.

[15] Department of Defense, Directive 3000.09, *Autonomy in Weapon Systems*, at http://www.esd.whs.mil/Portals/54/ Documents/DD/issuances/DODd/300009p.pdf.

[16] Most recently, in the U.S. commentary on the guiding principles adopted by the Group of Governmental Experts (GGE) on Convention on Certain Conventional Weapons (CCW) on lethal AWS, available at https://documents.unoda.org/wp-content/uploads/2020/09/20200901-United-States.pdf.

[17] National Security Commission on Artificial Intelligence, "Interim Report," p. 17.

[18] National Security Commission on Artificial Intelligence, "The Role of AI Technology in Pandemic Response and Preparedness: Recommended Investments and Initiatives," White Paper Series on Pandemic Response and Preparedness, No. 3, June 25, 2020, p. 27, available at https://drive.google.com/file/d/153DUHToD4zoM_GXe9MWGNKzend7TsI2o/view.

[19] Ibid, p. 8.

B. Maximizing security requires balancing pursuit of military and technological advantages with recognition of the dangers of international arms races, unintentional conflict escalation, harmful accidents, weapons proliferation, and nonstate malicious use of emerging technologies.

Geopolitical competitors to the United States, China in particular, continue to invest amounts equivalent to billions of dollars each year into AI research and development.[20] While the United States should respond with its own investments and seek to retain its commercial and military edge in AI, legislators and policymakers must also be careful not to let rational pursuit of technological advantage drive reckless innovation or sideline vital national interests.

As with nuclear weapons, AI seems destined to be a focal point for grand strategy in the years ahead. The field's strategic salience gives rise to a number of major risks. Experts have long warned of the possibility of costly AI arms races, which hold the potential to destabilize the international system and alter the balance of power to the detriment of U.S. interests and national security.[21] Beyond arms races, AI could invalidate long-standing assumptions in deterrence theory, including by altering perceptions regarding the credibility of a secure second-strike capability.[22] Artificial intelligence could also exacerbate security dilemmas: adversaries' AI capabilities are harder to assess than their conventional warfighting capabilities.[23] Lack of comprehensive international governance frameworks, meanwhile, heighten the risk of "flash wars"—rapid-onset unintentional conflicts that could arise due to escalatory spirals from automated or autonomous systems.[24] Premature deployment of AI systems could compromise safety as well.[25] As former U.S. Secretary of the Navy Richard Danzig writes, "the most

[20] Ashwin Acharya and Zachary Arnold, "Chinese Public AI R&D Spending: Provisional Findings," CSET Issue Brief, Center for Security and Emerging Technology, December 2019, p. 2, https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-2.pdf.

[21] For example, see: Congressional Research Service, "Artificial Intelligence and National Security," CRS Report, April 26, 2020, p. 17 and 30, https://crsreports.congress.gov/product/pdf/R/R45178/3; Chris Meserole, "Artificial Intelligence and the Security Dilemma," *Order From Chaos*, The Brookings Institution, November 6, 2018, https://www.brookings.edu/blog/order-from-chaos/2018/11/06/artificial-intelligence-and-the-security-dilemma/.

[22] Edward Geist and Andrew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?," RAND Corporation, 2018, pp. 10-11, available at https://www.rand.org/pubs/perspectives/PE296.html.

[23] Meserole, "Artificial Intelligence and the Security Dilemma."

[24] A RAND Corporation wargaming exercise found that the speed of autonomous systems led to inadvertent escalation. See: Yuna Huh Wong et al., "Deterrence in the Age of Thinking Machines," RAND Corporation, 2020, p. 52, available at https://www.rand.org/pubs/research_reports/RR2797.html.

[25] Stuart Armstrong, Nick Bostrom, and Carl Shulman, "Racing to the Precipice: A Model of Artificial Intelligence Development," Technical Report #2013-1, Future of Humanity Institute, Oxford University, October 2013, p. 1, https://www.fhi.ox.ac.uk/wp-content/uploads/Racing-to-the-precipice-a-model-of-artificial-intelligence-development.pdf.

reasonable expectation is that the introduction of complex, opaque, novel, and interactive technologies will produce accidents, emergent effects, and sabotage."[26]

Not only can a single-minded pursuit of technological advantage undermine national security and incur considerable costs; it can also erode American values if taken to an extreme. Open societies like the United States are constrained in their ability to maintain technological primacy due to the natural diffusion of capital, ideas, and expertise that international trade, science, and immigration engender. While trading off against this openness by implementing safeguards on student visas, foreign direct investment, intellectual property transfers, and other avenues for acquisition of American technologies is sometimes necessary, constructing walls that are too high or wide can jeopardize national competitiveness, fundamental liberties, and the integrity of the liberal order that has undergirded Western peace and prosperity for three-quarters of a century.

To ensure that U.S. leadership in AI is sustainable and consistent with American values, we recommend the following actions:

- **Create incentive structures within the Department of Defense that ensure safety is prioritized above expediency.** The Commission's stated view is that "[t]here is an ethical imperative to accelerate the fielding of safe, reliable, and secure AI systems that can be demonstrated to protect the American people, minimize operational dangers to U.S. service members, and make warfare more discriminating, which could reduce civilian casualties." We are concerned that political and military incentive structures could favor expediency over safety in decisions regarding the deployment of AI systems.[27] Incentive structures should reward red-teaming, bug bounties, and other practices that promote safety and reliability while holding accountable internal and external stakeholders who undermine the trustworthy application of AI in the national security enterprise. In order to increase AI safety and security, the United States should not make policy in isolation, but rather consider its effect on international arms markets and the military postures and behavior of state and non-state actors.
- **Expand mechanisms and protocols for rapid conflict de-escalation.** The sheer speed at which algorithmic escalation and warfare can occur—a mere fraction of a second in some cases—poses an immense challenge for conflict de-escalation. Limiting the damage that can ensue from unintentional conflict requires expanding lines of emergency communication between high-level U.S. officials and their counterparts in competitor nations; developing new international protocols for rapid crisis de-escalation, especially if autonomous weapons are involved; and requiring human oversight of AI

[26] Richard Danzig, "Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority," Center for a New American Security, May 30, 2018, p. 2, available at https://www.cnas.org/publications/reports/technology-roulette.
[27] National Security Commission on Artificial Intelligence, "Interim Report," p. 17.

systems exercising partial or full control over highly destructive weapons, nuclear in particular.[28]

- **Elaborate specific, well-defined use cases for proposed AI technologies prior to their development and use.** We concur with the Defense Innovation Board's contention that "DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use."[29] Deployment of AI systems, offensive ones in particular, can increase the probability that foreign governments perceive U.S. actions to be threatening, prompting them in some cases to adopt more aggressive force postures that exacerbate risk of unintended conflict escalation. Documenting and justifying specific use cases for novel AI systems prior to, during, and after their development allows the United States to identify such risks well in advance and plan accordingly. Demanding a clear rationale for investment into specific AI technologies also affords more realistic outlooks on battlefield and strategic implications, and reduces the likelihood of early obsolescence of key systems.
- **Incorporate AI safety standards into Department of Defense procurement processes.** National and international standard-setting bodies have begun to promulgate standards for transparency, trustworthiness, and other aspects of AI safety.[30] The Department of Defense should incorporate these standards into its procurement processes to ensure that commercially developed systems are safe, secure, and reliable. The Department should also designate liaisons for communication and collaboration with the National Institute of Standards and Technology and the American National Standards Institute to ensure that U.S. and international standards are stringent enough to meet high battlefield requirements.

C. <u>Maintaining robust diplomatic engagement with allied and nonallied nations alike is necessary to ensure that artificial intelligence enhances rather than undermines national security.</u>

While great power competition appears to be increasing and merits corresponding attention in U.S. strategic thinking, this trend does not by itself capture international political dynamics in their entirety. Across many geographic and issue areas, dynamics of "complex interdependence" continue to drive myriad critical aspects of international relations. Artificial

[28] Michael T. Klare, "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation," *Arms Control Today*, April 2020, https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation.

[29] Defense Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense," October 31, 2019, p. 8, https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

[30] For an example of pertinent activity in international standard-setting bodies, see: International Organization for Standardization/International Electrotechnical Commission Joint Technical Committee 1/SC 42, "Information Technology—Artificial Intelligence—Overview of Trustworthiness in Artificial Intelligence," ISO/IEC TR 24028:2020, May 2020, https://www.iso.org/standard/77608.html.

intelligence is no exception. Foreign nations supply the United States with skilled workers who comprise large portions of the U.S. scientific and technical workforce; critical components and assembly of AI-related hardware; and datasets used to train deep learning algorithms, *inter alia*. We commend the Commission for its recognition of this fact and its emphasis on maintaining close and productive relationships with U.S. allies in matters of AI research, development, and application.

Productive diplomatic engagement, however, must extend beyond allied nations if AI is to further American interests and enhance U.S. national security. The United States must work with allied *and* non-allied nations alike to establish international rules and norms for AI safety; limit proliferation of autonomous weapons and other AI technologies to terrorists, rogue states, criminal syndicates, and other malicious actors; implement safeguards against AI-related human rights abuses; and accelerate diffusion of beneficial AI technologies that promote economic development and sociopolitical stability, among other objectives. Achieving these goals requires working with geopolitical competitors to reach bilateral and multilateral agreements in service of common aims. If the United States neglects diplomatic strategy and engagement beyond allied nations, it risks missing out on opportunities to address collective security concerns and ceding influence to countries such as China and Russia.

To maintain productive diplomatic engagement with both allied and non-allied nations, we recommend the following actions:

- **Establish bilateral and multilateral dialogues on AI safety and other aspects of technological research, development, and application for which mutual understanding can enhance collective security.** Frequent track 1 and 1.5 dialogues can facilitate sharing of information, foster transparency and trust, spark new international collaborations, and open opportunities to gain intelligence on foreign AI-related activities. The Department of State and the AI research community should establish and support a sufficient number of such dialogues to ensure robust scientific diplomacy with allies, competitors, and other states alike. Abundant anecdotes of scientific diplomacy efforts during the Cold War[31] and in more recent times[32] attest to their efficacy and importance. Similar dialogues on AI could reduce geopolitical tensions while promoting its safe development.
- **Support nongovernmental fora that facilitate international dialogue on AI.** The United States should also sponsor delegations to multilateral and multi-stakeholder gatherings and institutions, such as the International Telecommunication Union's AI for Good Global Summit and the International Congress for the Governance of Artificial

[31] The Pugwash Conferences are among the most famous examples. See: Alison Kraft, Holger Nehring, and Carola Sachse, "The Pugwash Conferences and the Global Cold War: Scientists, Transnational Networks, and the Complexity of Nuclear Histories," *Journal of Cold War Studie*s 20.1 (2018): pp. 4-30, available at https://www.mitpressjournals.org/toc/jcws/20/1.

[32] The Agreement on Enhancing International Arctic Scientific Cooperation provides one recent example. See: Paul Berkman et al., "The Arctic Science Agreement propels science diplomacy," *Science* 6363 (November 2017): pp. 596-598, available at https://science.sciencemag.org/content/358/6363/596.

Intelligence, to ensure that American views are represented and help build favorable consensus around critical AI-related issues. U.S. participation in gatherings can facilitate mutual understanding beyond the auspices of government and building of scientific consensus. Such participation can also help ensure that gatherings are consistent with American values and strategic aims.

- **Extend technical advising on AI-related issues to countries receiving security and development assistance.** While the United States has dedicated considerable effort to coordinating with key allies in the development and application of AI technologies, including through the Pentagon's new AI Partnership for Defense, it has devoted far less attention and resources to building relationships with countries outside this select group. Neglecting AI development in and diffusion to such countries risks unsafe design and application of pertinent technologies, heightening security vulnerabilities and increasing the potential for catastrophic accidents. Extending technical advising on AI safety-related issues to recipients of security and development assistance can help mitigate these risks while simultaneously advancing U.S. diplomatic aims, including the negotiation and adoption of favorable international standards and agreements for AI. It can also increase situational awareness, affording policymakers the ability to respond more nimbly and ably to emerging developments. Finally, U.S. failure to provide technical advising may make countries more susceptible to influence from China and other geopolitical competitors, who may employ technology transfers and targeted investments to further their geostrategic agendas at the expense of the United States.

D. Formulating and implementing effective strategies around AI requires acknowledging the limits of foresight and designing institutions capable of adapting to unanticipated technological and geopolitical developments.

While competition with China and Russia presents salient challenges for the United States now and for the foreseeable future, geopolitical trends are not immutable. Historical evidence, including the failure of most analysts to predict the collapse of the Soviet Union and subsequent easing of U.S.-Russia tensions in the late 1980s and early 1990s, underscores how extrapolating current patterns of international relations and distributions of power can lead to erroneous predictions and misdirected action. The U.S. national security and foreign policy communities, then, should remain open to and plan for seemingly remote contingencies such as an easing of tensions vis-à-vis China or Russia, a falling out between the United States and a subset of its allies, or a major change in the structure of the international system.

Technological innovation is similarly recalcitrant to prediction. Since its inception as a distinct field of academic study in 1956, artificial intelligence has experienced alternating booms and busts, due in some cases to endogenous technical developments and in others to exogenous political and economic factors.[33] Over the decades, symbolic logic-based AI has given way to

[33] Bruce G. Buchanan, "A (Very) Brief History of Artificial Intelligence," *AI Magazine* 26.4 (2005), pp. 53-60, available at https://www.aaai.org/ojs/index.php/aimagazine/article/view/1848.

newer techniques such as deep learning, resulting in drastic changes in system architectures, resource requirements, and use cases. Similar dynamics could characterize future developments in the field. While many contemporary AI systems rely on large sets of labeled training data, for example, continued innovation could obviate this need. The high potential for AI to combine with other technologies in unforeseeable ways compounds the difficulty of prediction. The Commission is right, in this sense, to turn its attention to "AI's position within a constellation of emerging technologies that both enable and build upon one another."[34] We encourage the Commission to extend this line of analysis in its final report, though we caution against placing too much confidence in projections, scenario-building exercises, and other forecasting devices. While these tools are useful, the more fundamental task lies in devising "appropriate mechanisms" capable of adjusting to change as it occurs. Devising such appropriate mechanisms will help the Commission fulfill one of its core duties given by Congress.[35]

To build a solid yet flexible foundation for adaptation to future geopolitical and technological developments, we recommend the following actions:

- **Update AI-related directives and other guidance on a biannual basis at minimum.** Emerging technologies such as AI and machine learning often develop in unpredictable ways and at an exponentially increasing speed. In acknowledgment of this fact, any directives, principles, or other guidance related to the development or use of AI may become outdated quickly. Given a natural tendency to bureaucratic inertia, such guidance should have regular sunset and review cycles established from the outset. In addition, advisory bodies, such as the Defense Science Board, should be used to conduct as-needed reviews of Defense Department guidance to address relevant changes in AI safety, machine ethics, or other research. Such reviews should acknowledge emerging international AI norms and principles and seek to align national guidance where possible.
- **Establish an expert body to brief the federal government on emerging issues in AI ethics and responsibilities.** In its First Quarter Recommendations, the Commission proposed that Congress establish a body composed of experts from civil society, academia, and Federally Funded Research and Development Centers (FFRDCs) and convened under the National Institute of Standards and Technology and the National Science Foundation to brief the national security and intelligence communities on emerging issues in AI ethics and responsibilities.[36] We endorse this recommendation, and we further recommend increasing the frequency of the proposed body's briefings, such that they occur on a semiannual rather than annual basis. We also recommend making the body as transparent as possible, including by making recordings and

[34] National Security Commission on Artificial Intelligence, "Interim Report," p. 50.

[35] Sec. 1051(b)(2)(J) of P.L. 115-232 stipulates that the Commission should consider "the evolution of artificial intelligence and appropriate mechanisms for managing such technology related to national security and defense."

[36] National Security Commission on Artificial Intelligence, "First Quarter Recommendations," March 2020, p. 70, available at https://drive.google.com/file/d/1wkPh8Gb5drBrKBg6OhGu5oNaTEERbKss.

transcripts of its meetings, briefings, and other events readily available to the public, as the information conveyed will have significant value beyond the federal government.

- **Place greater emphasis on AI and other emerging technologies in macro-level geopolitical forecasts.** U.S. policymakers craft strategies based in part on their expectations about the future. Documents such as the National Intelligence Council's *Global Trends 2035*, which couched AI more in terms of its labor market than security implications, can play an important role in influencing these expectations.[37] In light of rapid technological innovation across multiple critical fields, the U.S. intelligence and national security communities should place greater emphasis on emerging technologies when engaging in forecasting exercises. The possible advent of artificial general intelligence in coming years or decades warrants particular attention, given its tremendous risk and uncertainty.[38]

E. Ensuring that artificially intelligent systems are safe, reliable, and ethical requires agencies that develop and implement such systems to be transparent and accountable.

Artificial intelligence presents a host of complex and unprecedented ethical challenges. Potential bias and discrimination, lack of transparency and explainability, and other issues threaten to undermine international human rights and humanitarian law, as well as cause catastrophic unintended harms. Legal and technical experts have questioned, for example, whether autonomous weapons systems can perform with sufficient predictability and reliability in the absence of meaningful human control.[39] Technical researchers, meanwhile, have revealed AI's susceptibility to adversarial attacks intended to cause system malfunctions. We believe that high thresholds are necessary to ensure that AI systems—and their operators—are in keeping with principles of ethical design and conduct, for both moral and pragmatic reasons.

To ensure that AI systems meet high thresholds for safety, reliability, and ethical conduct, we recommend the following actions:

- **Expand federal funding opportunities for AI safety and ethics research.** As part of its work on artificial intelligence, the National Science Foundation (NSF) has launched a cluster of National AI Institutes.[40] Making AI safety and ethics well-funded pillars of the

[37] National Intelligence Council, "Global Trends 2035: Paradox of Progress," January 2017, https://www.dni.gov/files/documents/nic/GT-Full-Report.pdf.

[38] On artificial general intelligence, see: Henry Kissinger, "How the Enlightenment Ends," *The Atlantic*, June 2018, https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

[39] International Committee of the Red Cross (ICRC), "Autonomy, artificial intelligence and robotics: Technical aspects of human control," August 2019, available at https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

[40] Emily K. Gibson, "New NSF AI Research Institutes to Push Forward the Frontiers of Artificial Intelligence," National Science Foundation, August 26, 2020,

National AI Institutes program would solidify research communities and accelerate innovation in these areas while helping to ensure that catastrophic accidents do not impede U.S. progress in the field as a whole. Along similar lines, NSF and other federal agencies should expand grantmaking and continue to launch new AI safety and ethics initiatives. Further, the Defense Department should establish a Multidisciplinary University Research Initiative (MURI) on AI safety, security, and robustness, as recommended by the Defense Innovation Board.[41] Supported research should encompass a wide range of scientific disciplines due to the ubiquity of human-machine interactions and the importance of the sociotechnical context of operation to AI performance. Areas of particular importance include, for example, psychological studies of "automation bias" and how sensory information from machine systems affects human decision-making regarding said systems, and how the operation of AI systems modifies the context in which they were designed to operate.

- **Provide rigorous training on AI safety and ethics across the U.S. national security enterprise.** We agree with the Commission's recommendation that an AI-ready national security workforce requires "training on the ethical and responsible development and fielding of AI at every level."[42] Training of critical personnel should recur on a frequent basis and be subject to periodic review by an independent advisory body composed of officials from relevant agencies and experts from the private sector, academia, and civil society. Such training is especially necessary for personnel in Offices of Inspectors General, who will necessarily review complicated use cases of AI systems. Federal agencies and departments should also create additional incentives for employees to obtain outside training and certifications in AI.
- **Ensure that Test, Evaluation, Verification, and Validation (TEVV) processes for autonomous systems account for tail risks.** In its Interim Report, the Commission asserted that "[t]here is a tension between fielding applications as quickly as possible and ensuring they perform reliably and safely. In finding the balance, we must not allow technical hurdles to serve as excuses for inaction."[43] While we recognize that AI holds the potential to confer tactical advantages and reduce American soldiers' exposure to harms, we believe that the higher risk of system failures associated with premature battlefield deployment of autonomous systems means that the Department of Defense should implement stringent TEVV processes that allow for expedited review only in highly specific and atypical circumstances. In all cases, these processes should account for low-likelihood, high-impact failures that would result in catastrophic outcomes, including, but not limited to, major conflict escalations and mass casualty events. Strengthening TEVV processes to account for such tail risk is particularly important when

https://beta.nsf.gov/science-matters/new-nsf-ai-research-institutes-push-forward-frontiers-artificial-intelligence.

[41] Defense Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense," Supporting Document, October 31, 2019, p. 44, https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.

[42] National Security Commission on Artificial Intelligence, "Interim Report," p. 36.

[43] Ibid., p. 33.

evaluating machine-to-machine and multi-agent interactions. As the Commission correctly notes, when individual AI systems are “combined in various ways in an enterprise to accomplish broader missions beyond the scope of any single system,” they are especially prone to various failure modes.[44] Unanticipated interactions between these complex systems could cause critical failures, as described in section A of this document, jeopardizing both mission outcomes and broader national security objectives. To help ensure that TEVV processes are rigorous and robust, they should also be subject to periodic external review by the Government Accountability Office, Department of Defense Office of the Inspector General, or a similar entity.

[44] National Security Commission on Artificial Intelligence, “Second Quarter Recommendations,” p. 145.