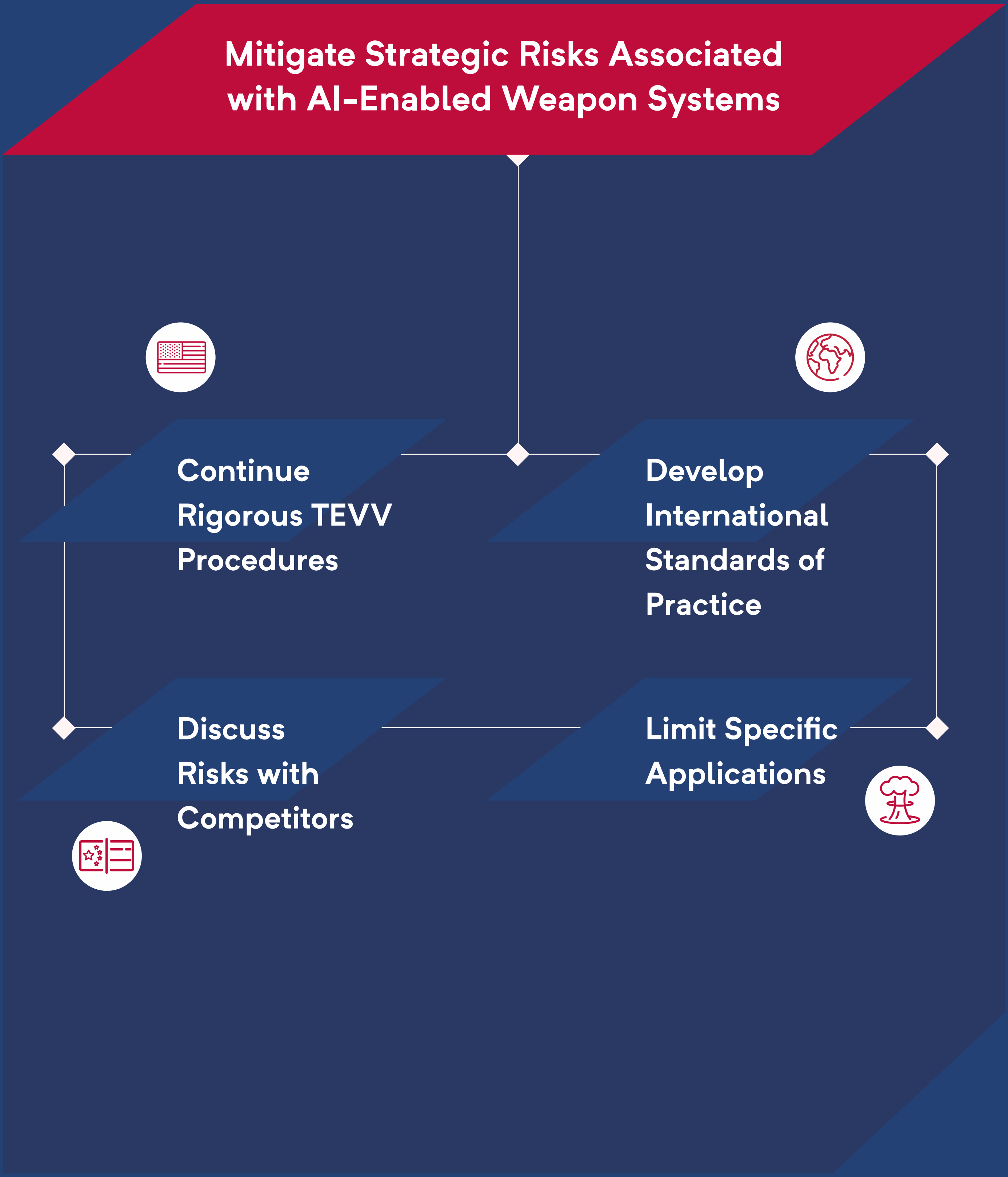
Chapter 4: Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare



Mitigate Strategic Risks Associated with AI-Enabled Weapon Systems

Continue Rigorous TEVV Procedures

Develop International Standards of Practice

Discuss Risks with Competitors

Limit Specific Applications

World military powers both large and small are pursuing artificial intelligence (AI)-enabled and autonomous weapon systems. Such systems have the potential to help commanders make faster, better, and more relevant decisions. They will enable weapon systems to be capable of levels of performance, speed, and discrimination that exceed human capabilities. And they will enable hitherto impossible complex tasks. If properly designed, tested, and used, they could improve compliance with International Humanitarian Law (IHL)[1] by reducing the risk of accidental engagements, decreasing civilian casualties, minimizing collateral infrastructure damage, and allowing for detailed auditing of the decisions and actions of operators and their command chains. Although U.S. weapons platforms have utilized autonomous functionalities for more than eight decades,[2] AI technologies have the potential to enable novel, sophisticated offensive and defensive autonomous capabilities.

The increasing use of AI technologies in weapon systems has generated important questions regarding whether such systems are lawful, safe, and ethical. Those critical of using AI technologies in weapons argue that states should negotiate limits or restrictions on such systems and their use. There is also concern that autonomous weapon systems may make conflict escalation more likely, and debate continues over what steps are needed to ensure that such systems minimize the risk of unintended military engagements or inadvertent and uncontrollable conflict escalation. Since 2014, the United Nations Convention on Certain Conventional Weapons (CCW) has held meetings among states parties to discuss the technological, military, legal, and ethical dimensions of "emerging technologies in the area of lethal autonomous weapon systems (LAWS)."[3] Specifically, it is examining whether autonomous technologies will be capable of complying with IHL and whether additional measures are necessary to ensure that humans maintain an appropriate degree of control over the use of force.

The Commission has consulted with civil society, academic organizations, and government agencies in studying the legal, ethical, and strategic questions that surround AI-enabled and autonomous weapon systems, including their potential military benefits and risks, possible ethical issues coming to the fore, international efforts to regulate them, and their compliance with IHL. The Commission offers the following four judgments to reflect its conclusions on these discussions.

Judgment 1: Provided their use is authorized by a human commander or operator, properly designed and tested AI-enabled and autonomous weapon systems have been and can continue to be used in ways which are consistent with IHL.

NSCAI Judgments Regarding AI-Enabled and Autonomous Weapon Systems

- Provided their use is authorized by a human commander or operator, properly designed and tested AI-enabled and autonomous weapon systems have been and can continue to be used in ways which are consistent with IHL.
- Existing DoD procedures are capable of ensuring that the United States will field safe and reliable AI-enabled and autonomous weapon systems and use them in a manner that is consistent with IHL.
- There is little evidence that U.S. competitors have equivalent rigorous procedures to ensure their AI-enabled and autonomous weapon systems will be responsibly designed and lawfully used.
- The Commission does not support a global prohibition of AI-enabled and autonomous weapon systems.

This judgment is grounded in several elements of IHL:

- **Distinction:** The principle of distinction holds that parties to an armed conflict must distinguish between civilians and combatants.[4] Weapons with increasingly accurate AI-enabled target recognition systems have the potential to reduce cases of target misidentification, the leading cause of inadvertent engagements during combat operations, and thus reduce civilian casualties and collateral infrastructure damage.[5]
- **Proportionality:** The principle of proportionality prohibits attacks which would cause incidental loss of civilian life excessive to the anticipated military advantage.[6] AI-enabled and autonomous weapon systems can and should also be designed to carry out operations in accordance with human judgments and directions regarding the proportionality of an attack. The moral reasoning involved in this calculus—weighing anticipated military advantage against potential civilian harm—remains the responsibility of a human commander.[7]
- **Accountability:** Ensuring accountability and command responsibility is essential to compliance with IHL. A human can and should be held accountable for the development, testing, use, and behavior of any autonomous weapon system, AI-enabled or otherwise. Autonomous weapon systems operate within the same general parameters as those used for human command and control systems, which are specifically designed to ensure accountability for actions and compliance with IHL. This is no different than for any other weapon system.[8]

The Commission endorses DoD's body of policy that states that human judgment must be involved in decisions to take human life in armed conflict. The kind of involvement necessary for humans to remain accountable for the use of autonomous weapon systems will vary depending on the time criticality of the situation as well as the operational context, circumstance, and type of weapon systems involved.[9] It is incumbent upon states to establish processes which ensure that appropriate levels of human judgment are relied

"... human judgment must be involved in decisions to take human life in armed conflict."

upon in the use of AI-enabled and autonomous weapon systems and that human operators of such systems remain accountable for the results of their employment.

Human accountability for the results of lethal engagements does not necessarily require human oversight of every step of an engagement process. Once a human authorizes an engagement against a target or group of targets, subsequent steps in the attack sequence can be completed autonomously without relinquishing human accountability. The exact number of steps in this sequence is dependent on the system's technical capabilities and the context and must consider factors such as the uncertainty associated with the system's behavior and potential outcomes, the magnitude of the threat, and the time available for action. For instance, an autonomous weapon system located in a rapidly changing environment, such as an urban setting, for an extended period, may require more frequent human authorization to ensure sufficient human accountability over autonomous actions than an equivalent system, operated for a similar amount of time, in a highly predictable and less populated environment—such as underwater or in space. This logic can and should be incorporated into the system's design, testing, and operational planning. Taking these factors into consideration, when feasible and deemed necessary operation designs should include points of required human guidance amid a sequence of automated actions. At such points, a human must review the system's status and authorize its next actions before the system's mission can continue. A blanket decision to compel every discrete step in an engagement involving lethal force to be subject to explicit authorization by a human is neither realistic nor desirable. Indeed, such a policy could instead spur commanders to use less precise, unguided weapon systems that might result in greater levels of collateral damage.

Judgment 2: Existing DoD procedures are capable of ensuring that the United States will field safe and reliable AI-enabled and autonomous weapon systems and use them in a manner that is consistent with IHL.

DoD's commitment to rigorous procedures for the development and use of autonomous weapon systems—as well as its commitment to strong AI ethical principles[10]—instills confidence that it will be able to field AI-enabled and autonomous weapon systems that are used lawfully. DoD has comprehensive processes for ensuring that the use of any weapon it fields is compliant with IHL and has a demonstrated commitment to operating within IHL, minimizing civilian casualties, and learning from its mistakes.[11] DoD has established a cross-department legal group, the DoD Law of War Working Group, to "develop and coordinate law of war initiatives and issues, such as analysis regarding the legality of new means or methods of warfare under consideration by DoD components."[12] This standing body is well positioned to examine implications for IHL as technology evolves over time. The International Committee on the Red Cross (ICRC) has lauded the strength and transparency of this system, listing the United States as one of eight countries that have "national mechanisms to review the legality of weapons and that have made the instruments setting up these mechanisms available to the ICRC."[13]

In addition to baseline legal review, the Department has taken special precautions for autonomous weapon systems to ensure these systems undergo sufficient test and evaluation, verification and validation (TEVV). In 2012, DoD added to an extensive list of guiding directives and instructions regarding weapons development within the Department by publishing DoD Directive (DoDD) 3000.09, Autonomy in Weapon systems, which establishes DoD policy for the development and use of autonomous weapon systems. It requires that all systems be designed "to allow commanders and operators to exercise appropriate levels of human judgment over the use of force" and requires senior DoD leaders to approve any autonomous weapon with lethal capabilities first when development begins, and again before fielding.[14] It also mandates any autonomous or semi-autonomous weapon that undergoes a revision to its operating state to undergo additional testing and evaluation. DoDD 3000.09 provides important definitions and baseline requirements for such systems and must be reviewed annually as technology evolves.[15] Chapter 7 of this report provides specific recommendations on how the United States should adapt its TEVV policies and capabilities to ensure it retains justified confidence in AI-enabled systems.[16]

"The U.S. commitment to IHL is longstanding, and AI-enabled and autonomous weapon systems will not change this commitment."

In addition, DoD's command and control procedures to authorize target selection and employment of munitions are rigorous and designed to ensure compliance with IHL. Operational commanders in the field are directly supported by lawyers embedded at multiple levels to advise on decisions about the use of force. The U.S. commitment to IHL is long-standing, and AI-enabled and autonomous weapon systems will not change this commitment.[17] These same principles will be ingrained into the design of those weapons, demonstrated in TEVV, and maintained by commanders overseeing their deployment. DoD's policy for autonomy in weapon systems and its adoption of ethical principles for AI in 2020 further highlight and reinforce this commitment.[18]

Judgment 3: There is little evidence that U.S. competitors have equivalent rigorous procedures to ensure their AI-enabled and autonomous weapon systems will be responsibly designed and lawfully used.

Battlefield success may become increasingly dependent on AI performance, and AI-enabled weapons are likely to proliferate given the open-source and dual-use nature of AI. This could cause pressure to mount on states to rapidly field new and untested systems and algorithms. Such pressures could also tilt designs toward systems that react more quickly, limiting the amount of time available for effective human oversight on engagement decisions. U.S. competitors, particularly Russia and China, likely do not have equivalent operational and targeting procedures to ensure the use of such systems is compliant with IHL and to preserve human accountability over the use of lethal force. Russia and China also have not published anything equivalent to DoDD 3000.09, outlining their policies and processes governing the acquisition, development, testing, and deployment of autonomous weapon systems. Unlike in the United States, in Russia and China these processes are secret, if they exist at all.

U.S. competitors have demonstrated that they are unlikely to adhere to the same ethical and legal standards in developing and utilizing AI-enabled weapon systems. Russia in particular has historically demonstrated a willingness to deploy risky and under-tested weapon systems, and it has deployed poorly performing unmanned ground vehicles

> **“A global treaty prohibiting the development, deployment, or use of AI-enabled and autonomous weapon systems is not currently in the interest of U.S. or international security ...”**

with limited autonomous functionalities in combat in Syria.[19] China is not only actively pursuing increased autonomous functionality across a range of military systems, but it is also currently exporting armed drones with autonomous functionalities to other nations. This includes systems such as the Blowfish A3, which Ziyan, the system's manufacturer, advertises as capable of conducting autonomous, lethal, targeted strikes.[20]

Judgment 4: The Commission does not support a global prohibition of AI-enabled and autonomous weapon systems.

A global treaty prohibiting the development, deployment, or use of AI-enabled and autonomous weapon systems is not currently in the interest of U.S. or international security and would be inadvisable to pursue for several reasons:

- First is the basic definitional problem. With respect to autonomous weapon systems, although the UN discussions about LAWS date back to 2014, states have yet to agree on a definition for them. This makes any treaty negotiation problematic, as it may be impossible to define the category of systems to be restricted in such a way that provides adequate clarity while not overly constraining existing U.S. military capabilities.
- Even if the definitional problem could be overcome, we judge that, at present, implementation of such an agreement would be impractical because compliance could not be verified. There is no feasible technical manner in which states could demonstrate to one another that specific weapon systems are or are not autonomous, or that they possess or lack certain capabilities. Doing so would require foreign inspectors to have short-notice access to the underlying code in weapon systems of concern. States are unlikely to agree to such an intrusive verification regime because revealing that information would create unacceptable risks to the security of their systems.
- Additionally, the effects of a prohibition agreement likely would run counter to U.S. strategic interests. Commitments from states such as Russia or China likely would be empty ones. Such an agreement would not serve the goal of putting political pressure on the states that are most likely to deploy autonomous weapon systems in unsafe and ethically concerning ways. Rather, the primary impact of an agreement would be to increase pressure on those countries that abide by international law, including the United States and its democratic allies and partners. Moreover, differing views on a prohibition among U.S. allies could deepen divisions among them on the employment of AI-enabled autonomous weapon systems. If U.S. allies joined an agreement while the United States did not, that divergence would likely hinder allied military interoperability.[21]

For these reasons, we believe the practical and strategic problems with a prohibition treaty outweigh potential benefits for the United States and its allies and partners, and therefore we support the current U.S. policy in opposition to such an agreement. However, this does not preclude other agreements or policies to address strategic risks associated with AI-enabled and autonomous weapon systems, or the future possibility of regulating specific types of technologies in AI-enabled and autonomous weapons technologies when such an agreement could be verifiable.

Recommendations to Mitigate Strategic Risks of AI.

While the Commission believes that properly designed, tested, and utilized AI-enabled and autonomous weapon systems will bring substantial military and even humanitarian benefit, the unchecked global use of such systems potentially risks unintended conflict escalation and crisis instability. The United States cannot assume that AI-enabled and autonomous weapon systems fielded by other countries will be developed, acquired, and fielded with the appropriate testing and verification to enable them to act as intended. Unintended escalations may occur for numerous reasons, including when systems fail to perform as intended, because of challenging and untested complexities of interaction between AI-enabled and autonomous weapon systems on the battlefield, and, more generally, as the result of machines or humans misperceiving signals or actions. AI-enabled systems will likely increase the pace and automation of warfare across the board, reducing the time and space available for de-escalatory measures. Beyond testing and robustness, we cannot assume that AI-enabled and autonomous weapons developed by other nations will be designed to behave in accordance with IHL.

Therefore, countries must take actions which focus on reducing risks associated with AI-enabled and autonomous weapon systems and encourage safety and compliance with IHL when discussing their development, deployment, and use. Such efforts should and must be led by the United States, which is uniquely situated to lead them given its technical expertise, military prowess, and clear and transparent policies and ethical principles governing the deployment and use of AI-enabled and autonomous weapon systems. The Commission presents the following five recommendations regarding actions the United States should take to mitigate risks associated with AI-enabled and autonomous weapon systems.

Strategic Risks Associated with AI-Enabled Weapons	NSCAI Recommended Actions	Objectives
	U.S. Actions	
Nation states could allow AI to authorize employment of key strategic weapon systems.	Clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons, and seek similar commitments from Russia and China.	Prevent unintended nuclear conflict due to AI-enabled launch authorization.
States cannot verify compliance with potential international agreements pertaining to AI-enabled weapons.	Pursue technical means to verify compliance with future arms control agreements pertaining to AI-enabled and autonomous weapon systems.	Enable effective verification of potential future agreements, which provides confidence systems are working as intended without revealing sensitive operational details.

Strategic Risks Associated with AI-Enabled Weapons	NSCAI Recommended Actions	Objectives
Global, unregulated proliferation of AI-enabled and autonomous weapons.	Fund research on technical means to prevent proliferation of AI-enabled and autonomous weapon systems.	Design and incorporate proliferation-resistant features into sophisticated AI-enabled and autonomous weapons, and potentially share them with Russia and China.
Poorly designed or improperly utilized AI-enabled weapons could behave unpredictably.	**U.S. Actions with Allies** Develop international standards of practice for the development and use of AI-enabled and autonomous weapon systems.	Set international norms guiding responsible development and use of AI-enabled and autonomous weapon systems.
AI-enabled systems could cause inadvertent conflict escalation.	**U.S. Actions with Russia and China** Discuss AI's impact on crisis stability in the existing U.S.-Russia Strategic Security Dialogue and create an equivalent meaningful dialogue with China.	Improve understanding of doctrine and develop confidence-building measures regarding use of AI-enabled and autonomous weapon systems.

Recommendation

Clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons, and seek similar commitments from Russia and China. The United States should make a clear, public statement that decisions to authorize nuclear weapons employment must only be made by humans, not by an AI-enabled or autonomous system, and should include such an affirmation in the DoD's next Nuclear Posture Review.[22] This would cement and highlight existing U.S. policy, which states that "[t]he decision to employ nuclear weapons requires the explicit authorization of the President of the United States."[23] It would also demonstrate a practical U.S. commitment to employing AI and autonomous functions in a responsible manner, limiting irresponsible capabilities, and preventing AI systems from escalating conflicts in dangerous ways. It could also have a stabilizing effect, as it would reduce competitors' fears of an AI-enabled, bolt-from-the-blue strike from the United States and could incentivize other countries to make equivalent pledges.

The United States should also actively press Russia and China, as well as other states that possess nuclear weapons, to issue similar statements. Although joint political commitments that only humans will authorize employment of nuclear weapons would not be verifiable, they could still be stabilizing, responding to a classic prisoner's dilemma: as long as countries have confidence that others are not building risky command and control structures that have the potential to inadvertently trigger massive nuclear escalation, they would have less incentive to develop such systems themselves.[24] While this norm is widely accepted in the United States, it is unclear if Russia and China share the same strategic

> **“… countries must take actions which focus on reducing risks associated with AI-enabled and autonomous weapon systems, and encourage safety and compliance with IHL when discussing their development, deployment, and use. Such efforts should and must be led by the United States …”**

concerns. Public reports indicate that Russia previously installed a “dead hand” system to automate nuclear launch authorization,[25] and China’s representatives in Track II dialogues with the United States have been hesitant to state that China would make an equivalent commitment. If neither Russia nor China is willing to agree to such a proposal, the United States should mount a strong international pressure campaign to condemn this decision and highlight how Russia and China refuse to commit to responsible military uses of AI.

Recommendation

Discuss AI’s impact on crisis stability in the existing U.S.-Russia Strategic Security Dialogue (SSD) and create an equivalent meaningful dialogue with China. The Departments of State and Defense should discuss AI’s impact on crisis stability within the existing U.S.-Russia SSD and create an equivalent meaningful dialogue with China. The SSD is an interagency bilateral dialogue focused on reducing misunderstandings and misperceptions on key strategic issues and threats, as well as reducing the likelihood of inadvertent escalation. Although the dialogue has traditionally focused on nuclear arms control and doctrine, it has recently been used to also discuss emerging technologies and space security.[26] The United States has no equivalent dialogue with China, as China has resisted U.S. attempts to establish one for nearly a decade. However, within the last year there has been increasing evidence that China is interested in formal talks with the United States concerning AI-enabled military systems.[27] This interest should be cultivated and leveraged into establishing a U.S.-China SSD that includes the relevant military, diplomatic, and security officials from both sides.

Given that the United States, Russia, and China are all aggressively pursuing AI-enabled capabilities, and that Russia and China are likely to field AI-enabled systems that have undergone less rigorous TEVV than comparable U.S. systems and may be unsafe or unreliable, it is crucial to improve mutual understanding of each other's military doctrines, including with respect to AI-enabled and autonomous systems. The United States should use this channel to highlight how deploying unsafe systems could risk inadvertent conflict escalation, emphasize the need to conduct rigorous TEVV, and discuss where each side sees risks of a conventional conflict rapidly escalating in order to better anticipate future responses in a crisis.

> **"… it is crucial to improve mutual understanding of each other's military doctrines, including with respect to AI-enabled and autonomous systems."**

These dialogues could also plant the seeds for a future, standing dialogue exclusively focused on establishing practical and concrete confidence building measures surrounding AI-enabled and autonomous weapon systems. For instance, the United States, Russia, and China could work to develop an "international autonomous incidents agreement," modeled after the 1972 Incidents at Sea Agreement, which would seek to define the "rules of the road" for behavior of autonomous military systems to create a more predictable operating environment and avoid accidents and miscalculations.[28] They could also agree to integrate "automated escalation tripwires" into systems that would prevent the automated escalation of conflict in specific scenarios without human intervention, to include nuclear weapons employment as noted above.

Recommendation

Work with allies to develop international standards of practice for the development, testing, and use of AI-enabled and autonomous weapon systems. The United States must work closely with its allies to develop standards of practice regarding how states should responsibly develop, test, and employ AI-enabled and autonomous weapon systems. This could build off of existing work, to include the 11 Guiding Principles agreed to by the LAWS Group of Governmental Experts (GGE) in 2019,[29] DoDD 3000.09, the DoD Ethical Principles for AI, and the NSCAI Key Considerations for Responsible Development and Fielding of AI.[30] As part of this effort, the DoD Law of War Working Group should meet regularly to review any future technical developments that pertain to autonomous weapon systems and IHL, and the tri-chaired Steering Committee on Emerging Technology (separately recommended by the Commission in Chapter 3 of this report) should advise on how such future technical developments impact policy and national defense.

The outputs of both groups should inform future DoD engagements with both allies and competitors on AI-enabled and autonomous weapon systems. Obtaining allied consensus regarding standards for the development, testing, and use of such systems will set important norms regarding these systems, help to ensure they are developed and used safely, and further highlight the commitment of the United States and its allies to ethical and responsible uses of AI. The United States should also use these consultations to highlight the ways in which AI will become a crucial part of future military operations and develop common frameworks guiding the appropriate and responsible use of AI-enabled and autonomous weapon systems on the battlefield. This should seek to incentivize allies to invest in the digital modernization of their own forces while also highlighting the risks to military interoperability should any ally agree to join a treaty prohibiting LAWS.

Recommendation

Pursue technical means to verify compliance with future arms control agreements pertaining to AI-enabled weapon systems. The United States should actively pursue the development of technologies and strategies that could enable effective and secure verification of future arms control agreements involving uses of AI technologies. Although arms control of AI-enabled weapon systems is currently technically unverifiable, effective verification will likely be necessary to achieve future legally binding restrictions on AI capabilities. DoD and the Department of Energy (DoE) should spearhead efforts to design and implement technologies which could provide other countries confidence that an AI-enabled and autonomous weapon system is working as intended without revealing sensitive operational details. For instance, it could examine ways for AI-enabled weapons platforms to produce authenticatable records of operation, which could be spot-checked via international challenge inspections if noncompliant activity is suspected. Technical creativity will be necessary to enable any future international restrictions on AI capabilities without revealing sensitive information.

Recommendation

Fund research on technical means to prevent proliferation of AI-enabled and autonomous weapon systems. Controlling the proliferation of AI-enabled and autonomous weapon systems poses significant challenges given the open-source, dual-use, and inherently transmissible nature of AI algorithms.[31] The proliferation of makeshift autonomous weapon systems which primarily utilize commercial components will be particularly difficult to control via regulation and will necessitate capable intelligence sharing and domestic law enforcement efforts to prevent their use by terrorists and other non-state actors. Regarding more sophisticated autonomous weapon systems, the United States should double down on efforts to design and incorporate proliferation-resistant features, such as standardized ways to prevent unauthorized users from utilizing such weapons, or reprogramming a system's functionality by changing key system parameters. DoD and DoE should fund technical research on such methods, and if appropriate, these methods could be shared with Russia and China, or potentially other countries, to prevent the proliferation or loss of control of certain AI-enabled autonomous weapon systems.[32]

This report does not contain a separate Blueprint for Action for Chapter 4. This is because given the importance of the topic, the Commission chose to detail its arguments, recommendations, and the specific actions required to implement them directly in this chapter. Additionally, further detail on how the United States should adapt its TEVV policies to maintain confidence in AI systems can be found in Chapter 7 and its associated Blueprint for Action, and recommendations on relevant changes to DoD organizational structure can be found in Chapter 3.

Chapter 4 - Endnotes

[1] IHL is also referred to as the law of armed conflict (LOAC) and the law of war.

[2] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W.W. Norton & Co. at 39 (April 24, 2018).

[3] *Background on Lethal Autonomous Weapon systems in the CCW*, United Nations (last accessed Jan. 11, 2021), https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument.

[4] *Distinction*, International Committee of the Red Cross (last accessed Jan. 15, 2021), https://casebook.icrc.org/glossary/distinction.

[5] There is room for improvement in reducing target misidentification in U.S. military operations. In the Afghanistan war, for example, a study indicated that about half of all civilian casualty incidents caused by U.S. forces resulted from target misidentification. The use of AI-enabled systems to make more accurate targeting decisions is perhaps the principal way in which the proper employment of AI could make warfare more humane. Larry Lewis, *Redefining Human Control: Lessons from the Battlefield for Autonomous Control*, CNA at 4 (March 2018), https://www.cna.org/cna_files/pdf/DRM-2017-U-016281-Final.pdf.

[6] *Proportionality*, International Committee of the Red Cross (last accessed Jan. 15, 2021), https://casebook.icrc.org/glossary/proportionality.

[7] See Paul Scharre, *Army of None: Autonomous Weapons and the Future of War,* W.W. Norton & Co. at 255-257 (2018).

[8] For a properly designed and tested autonomous system which correctly carries out the commander's intent, the commander is clearly accountable for the actions of that system. It is incumbent on states to properly design, test, and use such systems and also put in place rigorous procedures ensuring that any weapon use complies with IHL, including by ensuring individual accountability.

[9] The Commission believes DoD's existing formulation of "appropriate human judgment," discussed in the following Judgment, captures that necessary variation and ensures that any decision to employ lethal force begins with and is under the control of human judgment, and that a human ultimately will remain accountable for any decision to employ force.

[10] Press Release, U.S. Department of Defense, *DoD Adopts Ethical Principles for Artificial Intelligence* (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[11] DoDD 5000.01 requires any weapon fielded by DoD to undergo a legal review to ensure compliance with the Law of Armed Conflict (LOAC), adhering to the requirements set out in Article 36 of the Protocol Additional to the Geneva Conventions of 12 August 1949. DoDD 3000.09 and the DoD AI Ethics Principles build on top of this baseline. See *Department of Defense Directive 5000.01: The Defense Acquisition System*, U.S. Department of Defense at 9 (Sept. 9, 2020), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf?ver=2020-09-09-160307-310; *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977*, International Committee of the Red Cross (last accessed Jan. 5, 2021), https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/WebART/470-750045.

[12] *Department of Defense Directive No. 2311.01: DoD Law of War Program*, U.S. Department of Defense at 11 (July 2, 2020), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/231101p.pdf?ver=2020-07-02-143157-007.

[13] *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, International Committee of the Red Cross at 5, n. 8 (Jan. 2006), https://www.icrc.org/en/doc/assets/files/other/icrc_002_0902.pdf.

[14] *Department of Defense Directive 3000.09: Autonomy in Weapon systems*, U.S. Department of Defense at 2 (Nov. 21, 2012, incorp. change 1 May 8, 2017), https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. The weapons-review processes established in DoDD 3000.09 are designed specifically to ensure that any U.S. autonomous weapon system complies with IHL principles such as discrimination and proportionality while also maintaining appropriate levels of human judgment and ensuring accountability.

[15] *Department of Defense Instruction 5025.01: DoD Issuances Program* at 22 (Aug. 1, 2016, incorp. change 3 May 22, 2019), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/502501p.pdf?ver=2020-05-20-081854-657.

[16] See the Appendix of this report containing the abridged version of NSCAI's Key Considerations for Responsible Development & Fielding of AI. For additional details on the Commission's recommendation for future R&D needed to advance capabilities for Testing, Evaluation, Verification, and Validation of AI systems, see the section on "System Performance" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[17] The DoD Law of War manual serves as a detailed resource for all DoD personnel responsible for implementing the law of war and executing military operations. See *Department of Defense Law of War Manual*, U.S. Department of Defense (Dec. 2016), https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190.

[18] Press Release, U.S. Department of Defense, *DoD Adopts Ethical Principles for Artificial Intelligence* (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[19] David Axe, *Don't Panic, But Russia Is Training its Robot Tanks to Understand Human Speech*, Forbes (June 30, 2020), https://www.forbes.com/sites/davidaxe/2020/06/30/dont-panic-but-russia-is-training-its-robot-tanks-to-understand-human-speech/?sh=7373377914f2.

[20] Patrick Tucker, *SecDef: China Is Exporting Killer Robots to the Mideast*, Defense One (Nov. 5, 2019), https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/.

[21] The United States has expressed similar concerns with respect to treaties banning cluster munitions and nuclear weapons. See *Q&A: Convention on Cluster Munitions*, HRW (Nov. 6, 2010), https://www.hrw.org/news/2010/11/06/qa-convention-cluster-munitions#; Heather Williams, *What the Nuclear Ban Treaty Means for America's Allies*, War on the Rocks (Nov. 5, 2020), https://warontherocks.com/2020/11/what-the-nuclear-ban-treaty-means-for-americas-allies/. As of March 2021, no ally with which the United States has a mutual defense agreement has expressed support for a treaty banning LAWS.

[22] The Commission recognizes that AI should assist in some aspects of the nuclear command and control apparatus, such as early warning, early launch detection, and multi-sensor fusion to validate single sensor detections and potentially eliminate false detections.

[23] *Nuclear Matters Handbook 2020*, Office of the Deputy Assistant Secretary of Defense for Nuclear Matters at 18 (2020), https://fas.org/man/eprint/nmhb2020.pdf.

[24] There could be other reasons countries may delegate nuclear weapons launch authority to autonomous systems, particularly if leadership trusts machines to execute launch orders more than humans. A political agreement is unlikely to be able to address these concerns, although offering it would highlight how other nations are engaging in irresponsible and dangerous behavior.

[25] Michael Peck, *Russia's 'Dead Hand' Nuclear Doomsday Weapon is Back*, The National Interest (Dec. 12, 2018), https://nationalinterest.org/blog/buzz/russias-dead-hand-nuclear-doomsday-weapon-back-38492.

[26] Press Release, U.S. Department of State, *Deputy Secretary Sullivan's Participation in Strategic Security Dialogue with Russian Deputy Foreign Minister Sergey Ryabkov* (July 17, 2019), https://2017-2021.state.gov/deputy-secretary-sullivans-participation-in-strategic-security-dialogue-with-russian-deputy-foreign-minister-sergey-ryabkov/index.html; Press Release, U.S. Department of State, *The United States and Russia Hold Space Security Exchange* (July 28, 2020), https://2017-2021.state.gov/the-united-states-and-russia-hold-space-security-exchange/index.html.

[27] Over the last year, Chinese experts have participated actively in several Track II dialogues with U.S. experts on the safety of military AI systems, potentially signaling a desire for formal government-to-government communication on these issues.

Chapter 4 - Endnotes

[28] See Michael C. Horowitz and Paul Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, Center for a New American Security (Jan. 2021), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AI-and-International-Stability-Risks-and-Confidence-Building-Measures.pdf?mtime=20210112103229&focal=none.

[29] *Final Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapon systems, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effect*, CCW/MSP/2019/CRP.2/Rev.1, (Nov. 13-15, 2019), https://undocs.org/CCW/MSP/2019/9.

[30] See the Appendix of this report containing the abridged version of NSCAI's Key Considerations for Responsible Development & Fielding of AI. For additional details on the Commission's recommendation for future action on International collaboration and cooperation, see the section on "System Performance" in *Key Considerations for Responsible Development & Fielding of Artificial Intelligence: Extended Version*, NSCAI (2021) (on file with the Commission).

[31] See Chapter 14 of this report for additional information on the difficulty of using export controls to prevent the transfer of AI algorithms.

[32] Along these lines, the United States shared the technology for Permissive Action Links (PALs), which prevent the unauthorized arming of a nuclear weapon, with the Soviet Union in the 1970s. It is not clear if there is an equivalent technology to PALs for AI, one which would reduce the risk of unauthorized or accidental escalation by an AI system without simultaneously significantly increasing the military performance of that system. If equivalent technologies are developed, cooperation would have to be considered on a case-by-case basis.