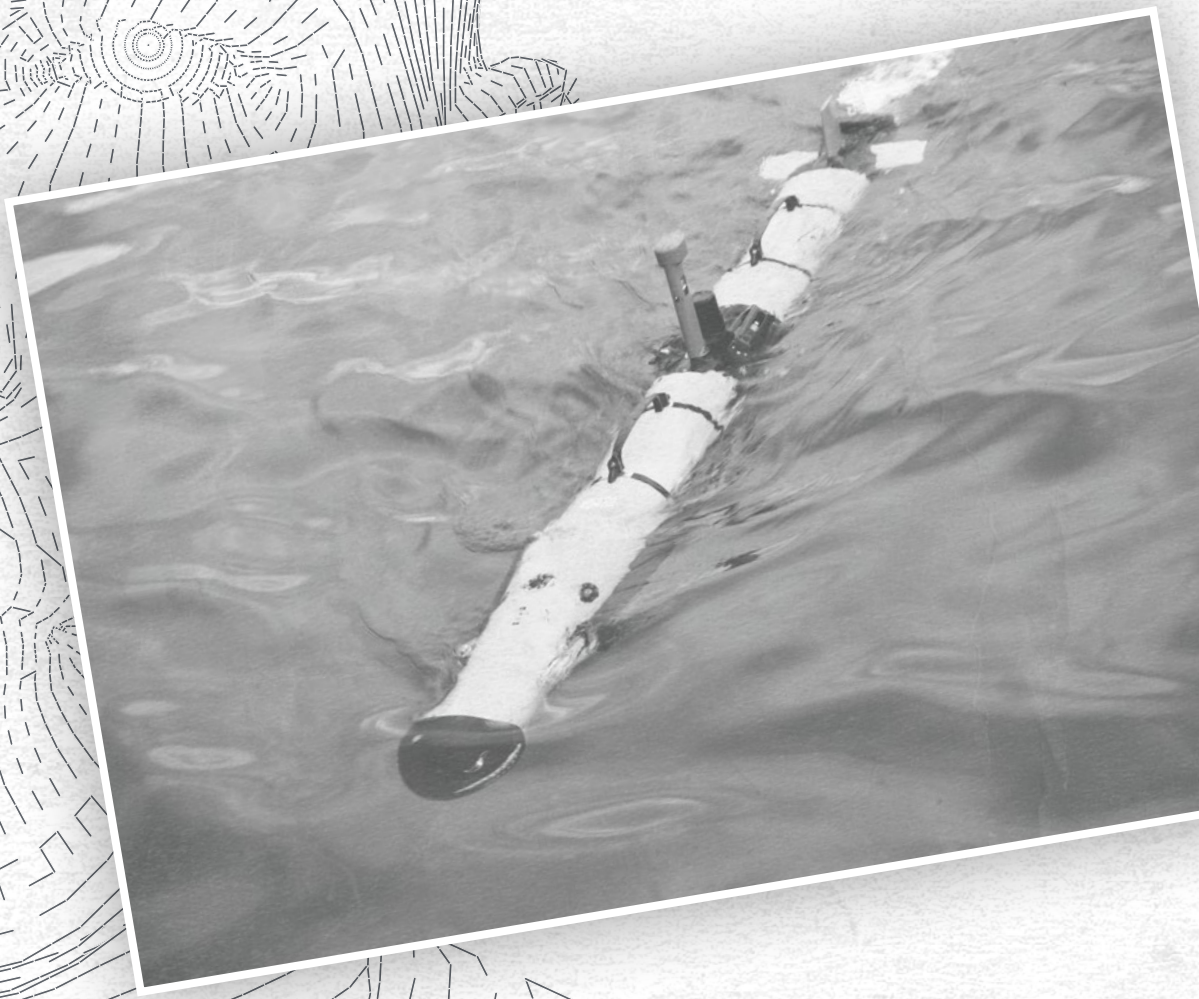
Texas National Security Review: Volume 6, Issue 2 (Spring 2023)
Print: ISSN 2576-1021 Online: ISSN 2576-1153

ARMS CONTROL FOR ARTIFICIAL INTELLIGENCE

Megan Lamberth
Paul Scharre



As AI continues to advance, some have voiced concerns about the dangers of AI-enabled weapons systems. This raises the question of how feasible it will be to control military use of AI. Megan Lamberth and Paul Scharre look at a number of characteristics that make AI difficult to control and lay out some concrete steps that could be taken today to increase the likelihood that future AI arms control regimes will be successful.

Militaries worldwide are working on how best to develop, integrate, and use AI in their weapons systems. While many of these systems are yet to be realized, breakthroughs in AI could have a significant impact on how militaries operate over time. Concern over military AI systems have led some activists to call for prohibitions or regulations on some AI-enabled weapons systems.[1]

Yet, AI has several characteristics that make it difficult to control. As a general-purpose enabling technology, AI is like electricity or the internal combustion engine and has countless nonmilitary or defense applications.[2] It differs from some military technologies because it is predominantly developed in the civilian sector by engineers in private industry or in research organizations. While the widespread availability of AI makes a complete ban on all military applications of AI unlikely, there may be an opportunity for the international community to work together to regulate or prohibit certain uses of military AI.

Throughout history, countries have sought restrictions or prohibitions for certain weapons or uses of weapons. The motivations for arms control can vary, as can its success. Evaluating historical cases of arms control shows that concrete steps taken today could increase the chances of successful AI arms control in the future. Policymakers can work to shape how AI technology is employed by militaries. Nations can also establish regular dialogue with allies and competitors on how AI might be used in warfare and what measures might be taken to reduce mutual risks.

The ubiquitous and democratized nature of AI makes arms control difficult but not impossible in all circumstances. While a total prohibition on military use of AI is unworkable, states could prohibit some applications of AI, provided that there was clarity on which uses were banned and that states had the ability to verify the compliance of other states. Verification, while challenging for any software-based military capability, could be achieved through a variety of possible methods: putting in place intrusive inspection regimes; regulating externally observable physical characteristics of AI-enabled systems (e.g., size, weight, payload) or autonomous behaviors; or restricting computing infrastructure (i.e., hardware). Any AI arms control would be challenging, but under the right conditions, it might be feasible in some cases. The right actions taken today can lay the groundwork for success in the future.

Types of Arms Control

Arms control encompasses a variety of actions and can occur at any stage of the development or use of a weapon. In this article, we define arms control as "agreements that states make to control the research, development, production, fielding, or employment of certain weapons, features of weapons, applications of weapons, or weapons

1 "Autonomous Weapons Open Letter: AI & Robotics Researchers," Future of Life Institute, Feb. 9, 2016, https://futureoflife.org/open-letter-autonomous-weapons/; "Lethal Autonomous Weapons Pledge," Future of Life Institute, June 6, 2018, https://futureoflife.org/lethal-autonomous-weapons-pledge/; Adam Satariano, "Will There Be a Ban on Killer Robots?" *New York Times*, Oct. 19, 2018, https://www.nytimes.com/2018/10/19/technology/artificial-intelligence-weapons.html; Tsuya Hisashi, "Can the Use of AI Weapons Be Banned?" *NHK*, April 18, 2019, https://www3.nhk.or.jp/nhkworld/en/news/backstories/441/; "Less Autonomy, More Humanity," Stop Killer Robots, accessed April 24, 2023, https://www.stopkillerrobots.org/; Matt McFarland, "Leading AI Researchers Vow to Not Develop Autonomous Weapons," *CNNMoney*, July 18, 2018, https://money.cnn.com/2018/07/18/technology/ai-autonomous-weapons/index.html; and Mary Wareham, "Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control," Human Rights Watch, August 2020, https://www.hrw.org/sites/default/files/media_2020/08/arms0820_web_0.pdf.

2 Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* 1, no. 3 (May 2018), https://doi.org/10.15781/T2639KP49.

delivery systems."[3]

Non-proliferation treaties, like the Nuclear Non-Proliferation Treaty, target the technology development phase and aim to prevent access to the underlying technology behind a certain weapon. Other arms control measures prohibit the development, production, and stockpiling of a weapon but allow access to the underlying technology. This includes bans on anti-personnel land mines and cluster munitions. Arms limitation treaties, such as the New START Treaty, allow for the production of certain weapons but attempt to limit the quantities that countries can possess.[4] Other measures regulate the use of a weapon in war, or in some cases, ban the use of a weapon entirely.

While some arms control measures are executed through legally binding agreements, there are numerous instances throughout history of successful non-legally binding agreements or even tacit cooperation without formal agreements. Over time, longstanding state practices may also evolve into customary international law, or "general practice accepted as law."[5]

Arms control is the exception, rather than the rule, when nations compete to develop and field weapons. Arms control requires coordination and trust among states — a difficult enough task in times of peace and an even harder one in times of conflict. States are often reluctant to agree to monitoring and verification measures that might enhance mutual transparency and enable states to verify others' compliance. As Andrew Coe and Jane Vaynman explain, an "important obstacle to arms control is the trade-off involved in monitoring: transparency is required to assure one side of the other's compliance with arms limits, but transparency might also reveal vulnerabilities that could be exploited by the first side in an arms race or war."[6] Despite the many obstacles to arms control, states have, in some circumstances, been able to successfully limit weapons development and use, even in war. The lessons from past historical successes and failures provide valuable insights for attempts to restrain militaries' pursuit of AI.

Why Arms Control Succeeds or Fails

Whether arms control succeeds or fails depends on both the desirability of arms control and its feasibility. The desirability of arms control is based on a country's calculation of a weapon's perceived military value versus its perceived horribleness, such as its inhumane effects on combatants, its indiscriminate nature, or its destabilizing effect on the international or political order. The feasibility of arms control depends on several factors: the ability of countries to achieve clarity on the level of desired restraint; the capacity for countries both to comply with an agreement and to verify other states' compliance; and the number of countries needed for an agreement to work. The likelihood that an arms control effort will be successful rises as desirability and feasibility increase.

Success in arms control exists on a spectrum. It entails limiting state behavior in the research, development, production, fielding, or use of a weapon. Measures that fail to restrain or regulate a behavior are considered unsuccessful. Most agreements fall somewhere in the middle. Even the most effective agreements, such as the Chemical Weapons Convention, have exceptions and violators. Other agreements may be successful for a period of time, but technology or changing political dynamics cause them to crumble. Take, for example, the Anti-Ballistic Missile Treaty and the Intermediate-Range Nuclear Forces Treaty. Yet, even partially successful arms control agreements can be effective at improving stability, minimizing civilian harm, and reducing combatant suffering.

3 Michael C. Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures," Center for a New American Security, Jan. 12, 2021, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures. Some definitions of arms control include post-conflict disarmament imposed by the victors on losing states, such as the Treaty of Versailles. For alternative definitions of arms control, see "Arms Control, Disarmament and Non-Proliferation in NATO," NATO, Feb. 27, 2023, https://www.nato.int/cps/en/natohq/topics_48895.htm; Thomas C. Schelling and Morton H. Halperin, *Strategy and Arms Control* (Washington, DC: Pergamon-Brassey's, 1985), 2; Robert R. Bowie, "Basic Requirements of Arms Control," *Daedalus* 89, no. 4 (Fall 1960): 708, http://www.jstor.org/stable/20026612; Hedley Bull, "Arms Control and World Order," *International Security* 1, no. 1 (Summer 1976): 3, https://doi.org/10.2307/2538573; Julian Schofield, "Arms Control Failure and the Balance of Power," *Canadian Journal of Political Science / Revue Canadienne de Science Politique* 33, no. 4 (December 2000): 748, http://www.jstor.org/stable/3232662; Coit D. Blacker and Gloria Duffy, *International Arms Control: Issues and Agreements* (Stanford, CA: Stanford University Press, 1984), 3; Lionel P. Fatton, "The Impotence of Conventional Arms Control: Why Do International Regimes Fail When They Are Most Needed?" *Contemporary Security Policy* 37, no. 2 (June 2016): 201, https://doi.org/10.1080/13523260.2016.1187952; and Henry A. Kissinger, "Arms Control, Inspection and Surprise Attack," *Foreign Affairs* 38, no. 4 (July 1960): 559, https://www.foreignaffairs.com/articles/1960-07-01/arms-control-inspection-and-surprise-attack.

4 "New START Treaty," U.S. Department of State, accessed April 24, 2023, https://www.state.gov/new-start/.

5 "Customary International Humanitarian Law," International Committee of the Red Cross, Oct. 29, 2010, https://www.icrc.org/en/document/customary-international-humanitarian-law-0.

6 Andrew J. Coe and Jane Vaynman, "Why Arms Control Is So Rare," *American Political Science Review* 114, no. 2 (May 2020): 342–55, https://doi.org/10.1017/S000305541900073X.

Desirability of Arms Control

States will be resistant to regulating a weapon with high military value — one that is effective, grants unique access, or provides a decisive battlefield advantage — even if the weapon has the capacity to cause substantial harm. Weighed against a weapon's military value is its perceived horribleness, that is, if the weapon is perceived to be inhumane, indiscriminate, destabilizing, or disruptive to the political or social order.

States have, at times, sought to restrict weapons or military systems that produce unnecessary suffering or superfluous injury.[7] For instance, bullets that leave glass shards in the body have a higher degree of horribleness — they cause excessive injury and the glass shards are undetectable by X-rays — and they do not provide a unique value to militaries. The perception of a weapon's horribleness may also be influenced by the mechanism of injury. Permanently blinding lasers, for example, are perceived to cause unnecessary suffering, which increases the desirability of controlling them.

States have tried to control less-discriminate weapons or military systems — those that cannot distinguish between civilian and combatant. This includes the early restrictions on aerial bombardment. These types of regulations are most successful when the weapon or behavior is banned altogether. Attempts to regulate the use of indiscriminate weapons by limiting their use to military targets and keeping them from civilian areas have not fared well in practice during wartime.

States may also desire arms control for weapons that are perceived to be disruptive or destabilizing. Political leaders may seek to ban a weapon that threatens their grip on power, such as the papal bans of the crossbow or early regulations on firearms. Weapons that are seen as destabilizing, such as anti-ballistic missile systems or space-based nuclear weapons, may have a higher desirability for arms control because they could provoke a costly arms race or could create perverse incentives for a first strike.

Reciprocity — the fear that another country might retaliate with a weapon or behavior in kind — is an essential factor in a state's desire for and compliance with arms control.[8] Before ratifying the 1925 Geneva Gas Protocol, the United Kingdom, France, and the Soviet Union declared that the agreement would cease to be binding if any one country failed to comply with it.[9] In the initial stages of World War II, Hitler refrained from ordering the bombing of British cities, not because of the international legal prohibitions in place against doing so but out of fear that Britain would retaliate in kind. Mutual restraint is achieved either by internal norms of appropriateness or fear of how an adversary might retaliate.[10] In a comprehensive study of law-of-war violations in 48 interstate wars from 1900 to 1991, James D. Morrow found that reciprocity was key to compliance.[11] Treaties were often a useful coordination mechanism for states to agree on whether and how to restrain their military operations, but violations on one side almost always led to reciprocal violations. In democracies, domestic institutions can create some "stickiness" that increases the likelihood of continuing to comply with treaties even when an opponent has violated them. Despite these pressures, Morrow concluded: "Unilateral restraint is rare."[12]

The best illustration of the dynamics of the desirability of arms control is the international community's response to nuclear weapons versus chemical weapons. The horribleness of nuclear weapons far outweighs the suffering that is caused by chemical weapons, yet global nuclear disarmament remains out of reach. Chemical weapons, on the other hand, have widely been denounced by the international community. Their occasional use has been by pariah states. The key difference lies in the military value of each weapon — nuclear weapons are uniquely politically effective. The result of this dynamic is that arms control is most often successful for weapons that are not especially valuable.

7 The use of "means and methods of warfare which are of a nature to cause superfluous injury or unnecessary suffering" is barred under customary international humanitarian law. "Rule 70. Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering," IHL database, International Committee of the Red Cross, accessed April 24, 2023, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule70.

8 Sean Watts, "Reciprocity and the Law of War," *Harvard International Law Journal* 50, no. 2 (Summer 2009): 365–434.

9 "Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare (Geneva Protocol)," June 17, 1925, U.S. Department of State, accessed April 24, 2023, https://2009-2017.state.gov/t/isn/4784.htm.

10 The "nuclear taboo" is a rare exception where international norms have played a role in the non-use of nuclear weapons, particularly against non-nuclear weapons states when reciprocal use is not an option. The nuclear taboo may be fraying in recent years, however. Nina Tannenwald, "The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use," *International Organization* 53, no. 3 (Summer 1999): 433–68, https://www.jstor.org/stable/2601286; and Nina Tannenwald, "The Vanishing Nuclear Taboo?" *Foreign Affairs* 97, no. 6 (November/December 2018), https://www.foreignaffairs.com/articles/world/2018-10-15/vanishing-nuclear-taboo.

11 James D. Morrow, "When Do States Follow the Laws of War?" *American Political Science Review* 101, no. 3 (August 2007): 559–72, https://www.jstor.org/stable/27644466.

12 Morrow, "When Do States Follow the Laws of War?" 570.

Feasibility of Arms Control

While the desirability of arms control has to do with the factors that motivate or discourage countries from pursuing arms control, feasibility involves whether mutual restraint is achievable. Whether or not arms control is feasible depends on several factors: the clarity of a regulation; the ability of states to comply with the terms of an agreement; the ability of states to verify other states' compliance; and the number of countries needed for an agreement to be successful.

For states to achieve arms control, they must agree on which weapons or uses of weapons are regulated and how they are to be regulated. Simplicity is a major advantage when crafting regulations. Complete bans on weapons have generally had better compliance in times of war than rules permitting use in some circumstances. Simplicity helps adversaries coordinate on restraint and aids in the normative value of stigmatizing a weapon. Absolute bans, like those on anti-personnel land mines, cluster munitions, and chemical and biological weapons, have been successful, in part, because the weapon is prohibited in all circumstances, not just in certain cases. Regulations that restrict use in some cases and not others, such as air-delivered weapons and submarine warfare, have historically been less likely to succeed.

And yet, simplicity is not always required. The United States and the Soviet Union / Russia engaged in multiple bilateral arms control agreements that had complicated rules for which weapons each state was permitted to build, including the Intermediate-Range Nuclear Forces Treaty, Anti-Ballistic Missile Treaty, SALT I, SALT II, SORT, START, and New START. However, these treaties had other advantages, such as only requiring two parties to reach an agreement and only limiting weapons development and production in peacetime, rather than wartime use.

Another significant factor affecting feasibility is a country's ability to actually comply with the conditions of an agreement. States have imperfect control over their armed forces, and regulations that states might accidentally violate may be harder to keep. Germany and Great Britain entered World War II seeking restraint on aerial bombing of cities, and mutual restraint held for a while. At first, both countries only bombed military targets, not populated areas. Restraint collapsed after German bombers accidentally flew off target in August 1940 and bombed central London by mistake. Britain retaliated by bombing Berlin, and Hitler responded by launching the London Blitz, after which all attempts at restraint were gone. Regulations that can easily be violated by chance are more difficult to maintain.

The feasibility of arms control is also affected by the number of countries needed for an agreement to succeed. The fewer countries necessary, the better. Throughout the Cold War, for example, the bipolar structure of the international system made arms control easier because only two superpowers needed to agree in order for arms control to succeed. Even for multilateral agreements, America and the Soviet Union could lead in developing an agreement, making it easier for other countries to follow. The United States and the Soviet Union established multiple arms control agreements, some bilateral and some multilateral. These agreements included the Seabed Treaty, Outer Space Treaty, Anti-Ballistic Missile Treaty, and others. The more diffuse a weapon, the harder it will be to control because more nations will be needed at the negotiating table.

A state's ability to verify whether other states are in compliance with an agreement is an important factor in making arms control succeed. Formal verification regimes have been used in some cases, particularly for weapons that can be developed in secret, such as nuclear or chemical weapons. The Chemical Weapons Convention and Nuclear Non-Proliferation Treaty include inspection measures to verify signatories' compliance. The Outer Space Treaty stipulates that states must allow others to view space launches and visit any facilities on the moon. Not all successful treaties have formal verification regimes, however. The bans on cluster munitions and anti-personnel mines do not require formal inspections but do compel states to be transparent with regard to eliminating their stockpile. In some cases, states rely on their own intelligence collection methods for verifying compliance, as was the case for the SALT I, SALT II, and Anti-Ballistic Missile treaties.

The legal status of an arms control agreement seems to have little impact on its ultimate success. Countries have violated legally binding treaties throughout history, particularly in times of war. This was the case with the use of poison gas in World War I. Informal, non-legally binding agreements have had success in the past. Take, for example, the Missile Technology Control Regime, which limits the export of certain classes of missiles. There are even some examples of a tacit understanding, where no formal agreement exists, restraining the use of a weapon, such as the restraint shown by America and the Soviet Union in deploying anti-satellite weapons or neutron bombs during the Cold War or Germany's unilateral recall of its sawback bayonet in World

War I. Rather than the legal status of a treaty, reciprocity is the driving factor that motivates states to comply. When states restrain their development, production, or use of a weapon, it is usually out of fear that competitors will respond to violations in kind. Formal agreements can be helpful, however, in coordinating state behavior and clarifying expectations.

Arms control is also path dependent: Regulations frequently build off prior successes of regulating related technologies. Ancient prohibitions on poison, for example, helped lead to modern bans on chemical and biological weapons. U.S.-Soviet arms control agreements on strategic weapons fostered additional agreements over time. The success of the humanitarian campaign to ban anti-personnel land mines likely made possible a humanitarian ban on cluster munitions.

Regulations on Evolving Technologies

States have frequently attempted to regulate new or rapidly evolving technologies, which present unique challenges that are especially relevant when it comes to AI. States may be uncertain about the military benefits and harms of emerging technologies and may err on the side of preserving the option to use them. When states desire restraint, they may fail to correctly predict the path of a technology's evolution, how it will be employed, and countermeasures that may be developed, causing states to craft regulations that are not practical or that fail to fully constrain harmful uses.

Nations embarked on a host of arms control agreements during the late-19th and early-20th centuries in an attempt to control industrial-age weapons. European powers signed arms control agreements in 1868, 1899, 1907, 1922, 1930, and 1936 regulating exploding or expanding bullets, poison gas, air-delivered weapons, submarines, and naval ships. None of these treaties attempted to stop proliferation of the underlying science and technology, such as chemistry or the internal combustion engine. Rather, they restricted the types of weapons that militaries could build or, in the case of aircraft and submarines, how the weapons could be used in war. Although European leaders correctly anticipated that many of these weapons, such as aircraft or poison gas, could cause great suffering in war, they failed to anticipate important details in how these technologies would evolve that, in some cases, hindered compliance.

Poison gas was outlawed prior to World War I, but only when released from projectiles. Germany's first large-scale use of chlorine at the Second Battle of Ypres was technically allowable because the gas was released from canisters. Achieving restraint with gas was further complicated by the fact that its relative military advantages and drawbacks were not known prior to World War I. Germany first employed gas in search of a war-winning weapon that would turn the tide at the front, perhaps motivated, in part, by French experiments with tear gas grenades early in the war.

Attempts to regulate submarines and aircraft similarly foundered on incorrect assumptions about how these technologies would evolve. The 1907 Hague Convention prohibited aerial bombardment against "undefended" cities, failing to anticipate the weakness of air defenses against bombing raids.[13] Conversely, maritime law required submarines to surface, give warning, and take the crew onboard before sinking merchant ships. Complying with these rules, which Germany initially tried to do in World War I, left submarines vulnerable to even lightly armed merchant ships.

In other cases, some weapons turned out to be not as problematic as states originally envisioned. Expanding bullets were banned in the 1899 Hague Convention, but today they are widely used by law enforcement and for civilian self-defense. (Expanding bullets are less likely to pass through a person and hit bystanders.)

More recently, in the case of the ban on blinding lasers, states have sought to sidestep the problem of predicting how the technology will evolve by adopting a ban on the intended use of the technology. Protocol IV of the Convention on Certain Conventional Weapons states: "It is prohibited to employ laser weapons specifically designed, as their sole combat function or as one of their combat functions, to cause permanent blindness to unenhanced vision."[14] The ban notably does not ban specific technical characteristics of a laser, such as its power, but instead focuses on its intended use. To date, the blinding laser ban has

13 "Convention (IV) Respecting the Laws and Customs of War on Land and Its Annex: Regulations Concerning the Laws and Customs of War on Land. The Hague, 18 October 1907," Article 25, International Committee of the Red Cross, accessed April 24, 2023, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule37_sectionc.

14 "Protocol on Blinding Laser Weapons (Protocol IV to the 1980 Convention), 13 October 1995," International Committee of the Red Cross, accessed April 24, 2023, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Treaty.xsp?action=openDocument&documentId=70D9427BB965B7CEC12563FB0061CFB2.

proven successful.

While preemptively banning new technologies is challenging in many ways, states have succeeded in imposing preemptive regulations of blinding lasers, biological weapons, using the environment as a weapon, placing nuclear weapons on the seabed or in space, and establishing military bases in Antarctica or on the moon. One factor that weighs in favor of preemptive regulations is that it may be easier, in some cases, to ban weapons that militaries have not yet integrated into their arsenals and are therefore not relying on for defense.

Lessons for Artificial Intelligence

AI will be difficult to control for three key reasons: It is a general-purpose technology; it is an emerging technology; and verifying the compliance of any AI-related agreement will pose unique challenges. This does not mean that AI is uncontrollable, however. While the obstacles are significant, arms control might be possible for some military AI applications. Even as countries pursue advantages in military AI, they should look for ways to mitigate the risks of military AI competition, including through arms control.

One factor that weighs in favor of preemptive regulations is that it may be easier, in some cases, to ban weapons that militaries have not yet integrated into their arsenals and are therefore not relying on for defense.

A General-Purpose Technology

As a general-purpose enabling technology, AI is more akin to electricity than it is to a discrete technology like submarines or blinding lasers, posing hurdles for arms control efforts. AI is a dual-use technology, with both civilian and military applications, and is likely to be widely available. The diffuse nature of AI makes a non-proliferation regime — one that would propose to "bottle up" AI and reduce its spread — unlikely to succeed. Additionally, because AI is so widespread, numerous actors would need to agree for any regulation to be successful.

The often fuzzy definition of "AI" also could complicate achieving clarity in any agreement. The definition of what constitutes "AI" is ambiguous and open to interpretation. Simply declaring, "No AI," lacks the clarity of outrightly banning all uses of gas because it may not be clear whether a technology qualifies as "AI." Additionally, because the field of AI is so vast and its uses are so wide ranging, banning all AI would be analogous to 19th-century states declaring "No industrialization." States did attempt to control industrial-age technologies, including submarines, aircraft, balloons, poison gas, and exploding or expanding bullets. But for countries to have agreed not to use any industrial-era technologies in warfare at all would have been impractical. It is also unclear, given the dual-use nature of civilian industrial infrastructure, if lines would or could have been drawn between civilian and military industrialization, even had countries desired them. Which military AI uses are acceptable or unacceptable could be ambiguous, and countries will need clarity and well-defined lines for any arms control efforts to be successful.

Historical cases of arms control during the industrial revolution serve as a useful guide because states did regulate, with varying degrees of success, specific applications of general-purpose industrial technologies, including the internal combustion engine (submarines and airplanes) and chemistry (exploding bullets and poison gas). These efforts sometimes failed, but this was not because states were unable to define what a submarine or airplane was or because states could not limit their civilian use. Rather, these measures failed due to how those weapons were specifically used in warfare. If the offense-defense balance between bombers and air defenses, or submarines and merchant ships had evolved differently, those weapons might have been controlled more effectively.

While a complete ban on all military AI applications is likely unattainable, history suggests that countries might be open to some limitations on specific applications. The challenge is determining for which specific applications of AI is arms control most desirable and feasible. Are there certain uses that are perceived as especially dangerous, destabilizing, or harmful? Scholars have already begun considering the impact of AI on nuclear stability, autonomous weapons, and cyber security,

and there will surely be other areas that merit serious consideration.[15] Ultimately, the desirability and feasibility of arms control for any specific military AI application may depend on how the technology is applied. An arms control agreement could be narrowly crafted to target specific instantiations of AI technology that are seen as particularly problematic, akin to countries' restricting bullets that are designed to explode inside the body, rather than exploding projectiles altogether.

An Emerging Technology

One challenge with anticipating which specific uses of military AI may merit further consideration for arms control is that it is unclear how AI will ultimately be used on the battlefield. This is a constant problem for emerging technologies, from historical examples such as airplanes and tanks to contemporary examples such as cyber tools and directed energy weapons. In the late-19th and early-20th centuries, countries struggled to control industrial-age technologies such as poison gas and submarines that were rapidly progressing.

The fact that AI is perceived by many militaries to be a "game-changing" technology could pose an obstacle to restraining its use. Militaries around the world are investing in AI and may be reluctant to restrict certain applications. The rhetoric surrounding AI — much of which may not match the actual investments militaries are making into AI — could itself hinder potential arms control efforts.

Additionally, perceptions that AI technology can yield superhuman capabilities, precision, reliability, or efficacy could diminish the belief that some applications of AI have the potential to be destabilizing or dangerous. These perceptions, even if they are unfounded, could impact a country's willingness to pursue AI arms control. As countries develop and field actual military AI applications, perceptions are likely to shift to align more closely with reality. But future military AI applications may be more difficult to regulate if they have already been integrated into a country's military forces or used on the battlefield.[16]

Verifying Compliance

The ability to verify compliance with any arms control agreement is essential for its long-term success. An agreement with clear language and buy-in from the necessary states could falter if states lack the means to verify others' compliance. AI complicates verification because — similar to other forms of software — an AI system's cognitive attributes are not easily externally observable. A "smart" weapon might look a lot like a "dumb" weapon of the same kind. An autonomous vehicle's sensors, which it uses to perceive its environment, may be visible, but the particular algorithm it uses might not be. This poses a challenge for any kind of arms control for military AI systems because mutual restraint relies on a state's ability to verify another state's compliance with an agreement. There are ways that countries could respond to this challenge, including by adopting intrusive inspections, restricting physical characteristics of AI-enabled systems, regulating observable behavior of AI systems, and restricting computing infrastructure.

Adopt Intrusive Inspections

An intrusive inspection regime could allow third-party observers access to a state's facilities and to certain military systems to verify that the state's software conforms to the stipulations of an arms control regime. Any potential inspection regime, however, would face the same transparency hurdles

15 Vincent Boulanin et al., "Artificial Intelligence, Strategic Stability and Nuclear Risk," Stockholm International Peace Research Institute, June 2020, https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk; Technology for Global Security, "AI and the Military: Forever Altering Strategic Stability," Technology for Global Security, Feb. 13, 2019, https://securityandtechnology.org/wp-content/uploads/2020/07/ai_and_the_military_forever_altering_strategic_stability__IST_research_paper.pdf; Forrest E. Morgan, et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World" (Santa Monica, CA: RAND Corporation, 2020), https://doi.org/10.7249/RR3139-1; Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," arXiv, December 2019, https://doi.org/10.48550/arXiv.1912.05291; Edward Geist and Andrew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?" RAND Corporation, 2018, https://doi.org/10.7249/PE296; Ben Buchanan, "A National Security Research Agenda for Cybersecurity and Artificial Intelligence," Center for Security and Emerging Technology, May 2020, https://doi.org/10.51593/2020CA001; Michael C. Horowitz, et al., "Policy Roundtable: Artificial Intelligence and International Security," *Texas National Security Review*, June 2, 2020, https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/; Melanie Sisson, et al., "The Militarization of Artificial Intelligence," Stanley Center for Peace and Security, August 2019, https://stanleycenter.org/publications/militarization-of-artificial-intelligence/; Giacomo Persi Paoli, et al., "Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making," United Nations Institute for Disarmament Research, 2020, https://www.unidir.org/publication/modernizing-arms-control; Andrew Imbrie and Elsa B. Kania, "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement," Center for Security and Emerging Technology, December 2019, https://doi.org/10.51593/20190051; Michael C. Horowitz, Lauren Kahn, and Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" *Orbis* 64, no. 4 (Fall 2020): 528–43; and Horowitz and Scharre, "AI and International Stability."

16 Rebecca Crootof, "Regulating New Weapons Technology," in *The Impact of Emerging Technologies on the Law of Armed Conflict*, ed. Eric Talbot Jensen and Ronald T.P. Alcala (New York: Oxford University Press, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3195980; and Rebecca Crootof and BJ Ard, "Structuring Techlaw," *Harvard Journal of Law and Technology* 34, no. 2 (Spring 2021), https://dx.doi.org/10.2139/ssrn.3664124.

that other weapons face — inspections risk exposing potential vulnerabilities in a state's weapons system to a competitor nation. This challenge could possibly be rectified in the future by privacy-preserving software verification, which could potentially verify the behavior of a piece of software without revealing private information.[17] Alternatively, countries could decide that the benefits to verification outweigh the risks of increased transparency. States have adopted intrusive inspection regimes in the past, such as inspections under the nuclear non-proliferation regime to verify civilian nuclear use.[18]

Another hurdle for inspections is that, if the difference between the permitted and banned capability lies in the software, a state could simply update its software after inspectors leave. Updating software is far quicker and easier than building another missile or nuclear enrichment facility. In the future, countries might be able to overcome this problem by adopting more advanced technical approaches. Potential options include continuously monitoring software to detect changes or embedding functionality into hardware, such that the capability is constrained by hardware, not software. Continuous monitoring would entail installing devices on military systems that would alert inspectors to any changes in software. Adopting such an approach requires further technological advancements, as well as states' commitment to continuous intrusive monitoring, rather than periodic inspections. It is also possible that such an approach, if implemented, could have unforeseen destabilizing effects in certain scenarios. For example, a software update to improve functionality on the eve of a conflict could trigger an alert that would lead other states to assume arms control noncompliance. Alternatively, regime-compliant code that should not be altered could be embedded into physical hardware, for example, through read-only memory or application-specific integrated circuits.[19] Intrusive inspection regimes will remain a weak option for verifying compliance unless states can confidently overcome the challenge of fast and scalable post-inspection updates to AI systems.

Restrict Externally Observable Physical Characteristics of AI-Enabled Systems

Instead of focusing on the cognitive abilities of an AI system, states could focus on the gross physical characteristics of a system that are easily observable and difficult to change, such as size, weight, power, endurance, payload, warhead, and so forth. This approach would allow states to adopt whatever cognitive characteristics (sensors, hardware, and software) they choose for a system. Arms control limitations would apply only to the gross physical characteristics of a vehicle or munition, even if the actual concern were motivated by the military capabilities that are enabled by AI. For example, if countries were concerned about swarms of anti-personnel small drones, instead of permitting only "dumb" small drones (which would be difficult to verify), they could prohibit all weaponized small drones, regardless of their cognitive abilities.[20] Countries have used similar approaches in the past — choosing to regulate gross physical characteristics (which were observable) as opposed to the actual payloads (which were states' actual concern, but more difficult to verify). Several Cold War-era treaties limited or banned certain classes of missiles, rather than only prohibiting arming them with nuclear weapons.[21] Limiting only nuclear-armed missiles would have permitted certain conventional missiles but would have been harder to verify.

Regulate Observable Behavior of AI Systems

Another option is for countries to center regulations on an AI system's observable behavior, such as how it operates under certain conditions. This would be analogous to the concept of "No cities" bombing restrictions, which did not prohibit bombers but instead regulated how they were employed. This approach could be used when dealing with physical manifestations of AI systems in which the outward behavior of the system is observable by other states. For example, countries could establish rules for how autonomous naval surface vessels

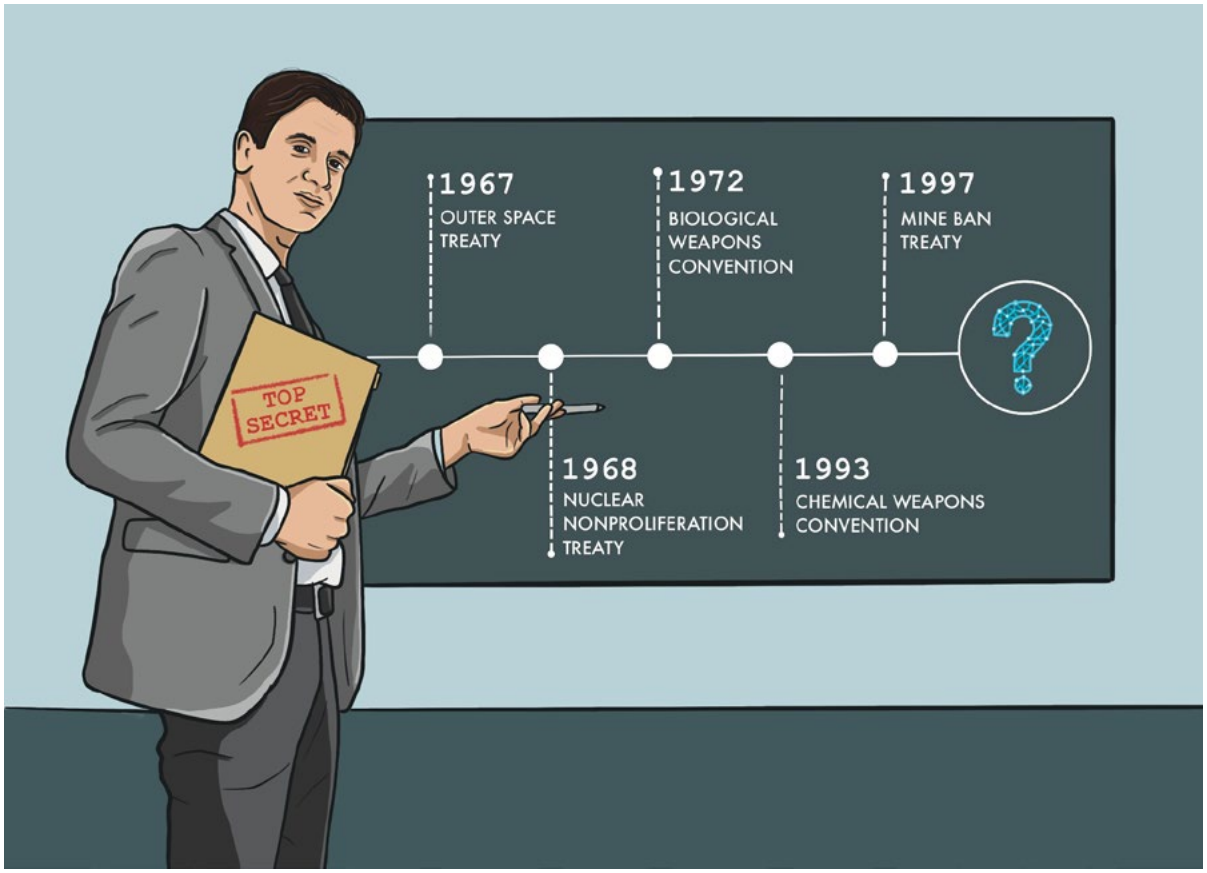
17 For more on privacy-preserving approaches for sharing information and verifying algorithms' behavior, see Andrew Trask, et al., "Beyond Privacy Trade-offs with Structured Transparency," arXiv, Dec. 15, 2020, https://doi.org/10.48550/arXiv.2012.08347; Joshua A. Kroll, et al., "Accountable Algorithms," *University of Pennsylvania Law Review* 165, no. 3 (2017), https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/; and Matthew Mittelsteadt, "AI Verification: Mechanisms to Ensure AI Arms Control Compliance," Center for Security and Emerging Technology, February 2021, https://doi.org/10.51593/20190020.

18 "More on Safeguards Agreements," International Atomic Energy Agency, accessed April 24, 2023, https://www.iaea.org/topics/safeguards-legal-framework/more-on-safeguards-agreements.

19 Mittelsteadt, "AI Verification," 18–24.

20 For an example of how such an approach might be implemented, see Ronald C. Arkin, et al., "A Path Towards Reasonable Autonomous Weapons Regulation: Experts Representing a Diversity of Views on Autonomous Weapons Systems Collaborate on a Realistic Policy Roadmap," *IEEE Spectrum*, Oct. 21, 2019, https://spectrum.ieee.org/a-path-towards-reasonable-autonomous-weapons-regulation.

21 For example, see the Intermediate-Nuclear Forces, SALT I, SALT II, START, SORT, and New START treaties. The Missile Technology Control Regime also regulates weapons of mass destruction-capable missiles.



ought to behave in proximity to other ships, even potentially adopting rules for how armed autonomous systems might clearly signal escalation of force to avoid inadvertent escalation in peacetime or crises. The particular algorithm used to program the behavior would be irrelevant — states could use different approaches. Similar to rules governing the behavior of human combatants, the regulation would govern how the AI system behaved, not its internal logic. This approach would not be effective, however, for military AI applications that are not observable. For instance, restrictions on the role of AI in nuclear command-and-control would likely not be observable by an adversary. This approach is also limited because the behavior of a system could be quickly modified through a software update, thereby undermining trust and verifiability.

Restrict Computing Infrastructure

States could focus regulations on elements of AI hardware that can be observed or controlled. AI systems rely on chips for computation, so countries could potentially restrict or control specialized AI chips through a non-proliferation regime (particularly if these chips were essential for the prohibited military capability).[22] Countries could also conceivably choose to restrict large-scale computing resources, also known as "compute," if they are observable or could be tracked. Large AI models such as GPT-4 are becoming increasingly general purpose and are able to execute a diverse range of tasks. These highly capable AI systems are inherently dual use. Embedded in their more general-purpose functionality by default are security-relevant capabilities, such as the capability to empower actors to launch cyber, chemical, or biological attacks.[23] Compute governance entails controlling the use of compute throughout the production lifecycle of an AI model, from chip manufacturing through model training and use. Current trends in AI suggest that restricting access to large-scale compute could be a particularly effective approach for denying access to the most cutting-edge AI capabilities.

22 Saif M. Khan, "U.S. Semiconductor Exports to China: Current Policies and Trends," Center for Security and Emerging Technology, October 2020, https://doi.org/10.51593/20200039.

23 "GPT-4 Technical Report," OpenAI, 2023, https://cdn.openai.com/papers/gpt-4.pdf; and "GPT-4 System Card," OpenAI, March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Frontier AI research labs have invested heavily in large-scale compute for machine learning in recent years. The leading AI research models are trained on thousands of specialized AI chips, such as graphics processing units, or GPUs, running for weeks at a time.[24] The amount of compute used in training frontier AI models grew ten billionfold from 2010 to 2022, doubling every six months (for the largest models, it doubled every 10 months).[25] This rate of growth is much faster than the 24-month doubling period under Moore's Law and is faster than the current rate of hardware improvements, which have been doubling every two and a half years.[26] To achieve this compute growth, the costs of large-scale training runs are skyrocketing. Independent estimates place the costs for training the largest models at least in the millions of dollars — possibly in the tens of millions — for the final training run.[27] Rising costs are consolidating progress at the cutting-edge of AI research in a handful of deep-pocketed tech companies, shutting out academic researchers from training the most compute-intensive models.[28] While this poses problems for the health and diversity of the AI research community, rising barriers to entry present an opportunity for controlling access to these AI capabilities.

While this poses problems for the health and diversity of the AI research community, rising barriers to entry present an opportunity for controlling access to these AI capabilities.

Because chips are a controllable physical resource, access to compute-intensive AI capabilities can be restricted by controlling access to high-end AI chips. Restricting access to large-scale compute is a particularly attractive approach because it would work even for a somewhat "leaky" regime, since prohibited actors must assemble large amounts of compute to be effective. In fact, the U.S. government took precisely this approach in October 2022 when it issued sweeping export controls on AI chips destined for China, even when these chips were manufactured outside of the United States.[29] If successful, U.S. export controls on high-end chips will effectively lock China out of the most advanced AI capabilities.

The feasibility of arms control aimed at AI hardware will depend heavily on the extent to which chip fabrication infrastructure is democratized globally or is concentrated in the hands of a few actors. While today's semiconductor supply chains are highly globalized, they also contain

24 For example, see Aakanksha Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways*, arXiv.org, April 19, 2022, https://doi.org/10.48550/arXiv.2204.02311. For a more comprehensive assessment of trends in data, compute, and model size in AI research, see Jaime Sevilla et al., "Parameter, Compute and Data Trends in Machine Learning," 2021, https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXN-fn3_YsT6sHXtCU0q7OIPuc4/.

25 "AI and Compute," OpenAI, May 16, 2018, https://openai.com/blog/ai-and-compute/; and Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, arXiv.org, March 9, 2022, https://doi.org/10.48550/arXiv.2202.05924.

26 Gordon E. Moore, "Cramming More Components Onto Integrated Circuits," *Electronics* 38, no. 8 (April 19, 1965), https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf; and Marius Hobbhahn and Tamay Besiroglu, "Trends in GPU price-performance," Epoch, June 27, 2022, https://epochai.org/blog/trends-in-gpu-price-performance.

27 Ben Cottier, "Trends in the Dollar Training Cost of Machine Learning Systems," Epoch, Jan. 31, 2023, https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems; Andrew J. Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?" Center for Security and Emerging Technology, Jan. 2022, 9, https://doi.org/10.51593/2021CA009; Sharir et al., *The Cost of Training NLP Models*; and Saif M. Khan and Alexander Mann, *AI Chips: What They Are and Why They Matter*, Center for Security and Emerging Technology, April 2020, 26, https://doi.org/10.51593/20190014. Other experts have estimated higher costs — up to tens of millions of dollars — for training some AI models. See Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, 22; Lennart Heim, "Estimating PaLM's Training Cost," *blog.heim.xyz*, April 5, 2022, https://blog.heim.xyz/palm-training-cost/; Dan H, "How Much Did AlphaGo Zero Cost?" Dansplaining, updated June 2020, https://www.yuzeh.com/data/agz-cost.html; and Ryan Carey, "Interpreting AI Compute Trends," AI Impacts, July 10, 2018, https://aiimpacts.org/interpreting-ai-compute-trends/. OpenAI CEO Sam Altman stated in April 2023 that the training for GPT-4 cost over $100 million. However, it is not clear whether that cost figure was for the final training run or for the total cost, including experiments prior to the final training run, and whether the cost included researcher salaries. Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

28 Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence, Aug. 18, 2021, https://arxiv.org/pdf/2108.07258.pdf; Rodney Brooks, "A Better Lesson," Rodney Brooks, March 19, 2019, https://rodneybrooks.com/a-better-lesson/; and Kevin Vu, "Compute Goes Brrr: Revisiting Sutton's Bitter Lesson for Artificial Intelligence," DZone.com, March 11, 2021, https://dzone.com/articles/compute-goes-brrr-revisiting-suttons-bitter-lesson.

29 "Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification," Federal Register, document no. 2022-21658, Oct. 13, 2022, https://www.federalregister.gov/public-inspection/2022-21658/additional-export-controls-certain-advanced-computing-and-semiconductor-manufacturing-items.

key chokepoints. These bottlenecks give a few countries the ability to control access to AI hardware. U.S. export controls on advanced AI chips to China are possible because of U.S. companies' dominance in semiconductor manufacturing equipment. American controls prohibit the use of U.S.-made equipment to manufacture high-end AI chips destined for China, even when those chips are made outside of the United States. Countries that dominate future chip supply chain chokepoints could potentially employ similar measures to control access to AI hardware.[30]

The future of semiconductor supply chains is highly uncertain, however. Supply chain shocks and geopolitical competition have accelerated state intervention in the global semiconductor market and have caused significant uncertainties in how the market will evolve. Recent U.S. export controls are only the latest state intervention in global semiconductor markets, and it will take time for the second- and third-order effects of these controls to play out. Some current trends seem to be pointing toward greater concentration of hardware supply chains, while other trends point toward greater democratization. One factor contributing to greater concentration in the industry is the high cost of semiconductor fabrication plants, or "fabs." On the other hand, China and the United States are both working hard to increase indigenous chip production for national security reasons, in both cases pushing against natural market consolidation as they spend government resources to subsidize new fabs. U.S. export controls themselves have the side effect of creating financial incentives for the private sector to circumvent U.S. controls by redesigning their chip manufacturing equipment to not rely on U.S. technology in order to sell to China's market. China imports over $400 billion a year in chips.[31] While U.S. export controls currently affect only an estimated 1 percent of the Chinese chip market, the market for banned chips is likely to grow if U.S. controls stay in place (as U.S. officials have said they will) and today's leading-edge chips become tomorrow's legacy chips.[32] Powerful market and nonmarket forces are impacting the global semiconductor industry, and the long-term effects of these forces on supply chains remains unclear.

Trends in improved algorithmic efficiency could also undermine the effectiveness of controlling compute in order to control AI capabilities. While the amount of compute used for training cutting-edge AI research models has grown over time, once a breakthrough is achieved, algorithmic improvements reduce the amount of compute required to achieve that same level of performance. For example, the amount of compute required to achieve the same level of image classification performance on ImageNet, an image recognition database, halved every nine months from 2012 to 2021.[33] Improvements in algorithmic efficiency can rapidly democratize the availability of AI models by lowering the amount of computing hardware needed to train models, making them more accessible.

One final challenge of using compute to control access to AI capabilities is the fundamental asymmetry in the compute resources required for training AI models relative to using them, a process known as "inference." Training an AI model on data is very compute-intensive, requiring massive amounts of data and compute for the largest models. Once a model is trained, however, the process of using the trained model to perform a task, such as generating text, classifying an image, or recognizing a face, generally uses much less compute. This means that the most effective point in the AI development pipeline for controlling access via compute is at the training stage. Limiting which actors have access to large amounts of compute — and regulating the behavior of those that do — could be an effective method of restricting access to AI capabilities.[34] However, once a model has been trained, compute becomes a far less effective point of control. Trained models can proliferate rapidly and have much lower barriers to entry for use in terms of the data, compute, and human capital requirements relative to training new models.

Trained models can proliferate if actors intentionally release open-source versions or if the models are stolen or leaked. At present, breakthrough AI models are quickly replicated with open-source

30 Saif M. Khan, Alexander Mann, and Dahlia Peterson, *The Semiconductor Supply Chain: Assessing National Competitiveness*, Center for Security and Emerging Technology, January 2021, 26, https://doi.org/10.51593/20190016.

31 "US Sanctions Help China Supercharge Its Chipmaking Industry," *Bloomberg*, June 20, 2022, https://www.bloomberg.com/news/articles/2022-06-20/us-sanctions-helped-china-supercharge-its-chipmaking-industry.

32 Alan Estevez and Martijn Rasser, "A Conversation with Under Secretary of Commerce Alan F. Estevez," Transcript of speech delivered at the Center for a New American Security, Washington, DC, Oct. 27, 2022, https://www.cnas.org/publications/transcript/a-conversation-with-under-secretary-of-commerce-alan-f-estevez.

33 Ege Erdil and Tamay Besiroglu, *Algorithmic Progress in Computer Vision*, arXiv.org, Dec. 16, 2022, https://doi.org/10.48550/arXiv.2212.05153; and Ege Erdil and Tamay Besiroglu, "Revisiting Algorithmic Progress," Epoch, https://epochai.org/blog/revisiting-algorithmic-progress; and "AI and Efficiency," OpenAI, May 5, 2020, https://openai.com/blog/ai-and-efficiency/.

34 Yonadav Shavit, "What Does it take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring," arXiv, March 20, 2023, https://doi.org/10.48550/arXiv.2303.11341.

versions. Open-source equivalents of the generative AI models GPT-3 and DALL·E were released after 14 and 15 months, respectively.[35] Once models are publicly available, they proliferate rapidly. Another way for trained models to proliferate is for the model to be leaked or stolen. Meta's AI language model LLaMA leaked on 4chan, circumventing Meta's attempts to limit access to it.[36] Once a trained model has been publicly released, compute may cease to be an effective point of control, since compute requirements for inference on trained models are relatively low. Trained models can also be fine tuned for specific uses through additional training, but without having to redo the costly initial training. Export controls on chips may need to be paired with export controls on trained models above a certain compute threshold, if they are to be effective in restricting access to high-end AI capabilities.

Looking Ahead

The current challenges of controlling AI-enabled military capabilities most closely resemble the militarization of industrial-age technology around the turn of the 20th century, when countries attempted to control an array of dangerous new weapons. Leading military powers met over 15 times to discuss a variety of arms control initiatives from 1868 to 1938.[37] The scale of this diplomatic activity conveys the level of persistence and patience needed to achieve even modest results in arms control. Policymakers, scholars, and civil-society advocates can take a number of steps today to begin to lay the foundations for future AI arms control. These include increasing dialogue about the risks that AI poses and the potential arms control measures that could be imposed, creating norms for the appropriate uses of military AI, tackling "low-hanging fruit" to build patterns of state cooperation on military AI, and shaping the development of AI technology itself to make it more controllable in the future. While none of these steps guarantees that future AI arms control efforts will be successful, they may increase the probability of success.

Increasing Dialogue

Increasing dialogue at all levels to better understand how AI might be used on the battlefield could help illuminate arms control measures that may be both desirable and feasible. Academic conferences, Track II academic-to-academic exchanges, bilateral and multilateral meetings, and discussions in international forums are all valuable for helping to advance mutual understanding among international parties.[38] These dialogues should also include AI scientists and engineers, to ensure that conversations are grounded in technical realities. Because AI technology is being driven by the commercial sector, they should also include major tech companies and organizations that have been engaged in norm development on AI, such as Microsoft, Google, OpenAI, Baidu, Tencent, and the Beijing Academy of Artificial Intelligence.[39] It is important that these discussions not engage in "ethics-washing," legitimizing improper uses of AI technology, such as for human rights abuses.[40] It is all too easy for institutions to put out well-meaning principles and statements about responsible AI. These statements must be paired with actions that demonstrate follow-through in using AI responsibly. Shaping norms for AI use as these norms are

35 Nathan Benaich and Ian Hogarth, *State of AI Report 2022*, [presentation], Oct. 11, 2022, https://www.stateof.ai/, slides 34–36.

36 James Vincent, "Meta's Powerful AI Language Model Has Leaked Online – What Happens Now?" *The Verge*, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

37 Leading military powers at the time met to discuss arms control in 1868, 1874, 1899, 1907, 1909, 1919, 1921, 1922, 1923, 1925, 1927, 1930, 1932, 1933, 1934, 1935, 1936, and 1938.

38 For example, the Responsible AI in the Military Domain (REAIM) summit held at The Hague in 2023. "About REAIM 2023," Government of the Netherlands, https://www.government.nl/ministries/ministry-of-foreign-affairs/activiteiten/reaim/about-reaim-2023. For more on the potential for multilateral dialogues, see Horowitz and Scharre, "AI and International Stability."

39 Rebecca Arcesati, "Lofty Principles, Conflicting Incentives: AI Ethics and Governance in China," Mercator Institute for China Studies, June 24, 2021, https://merics.org/en/report/lofty-principles-conflicting-incentives-ai-ethics-and-governance-china; Matt Sheehan, "China's New AI Governance Initiatives Shouldn't Be Ignored," Carnegie Endowment for International Peace, Jan. 4, 2022, https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127; Jessica Cussins Newman, *AI Principles in Context*, Asia Society, Aug. 20, 2020, https://asiasociety.org/sites/default/files/inline-files/Cussins_Principles_Final.pdf; "司晓：打造伦理"方舟"，让人工智能可知、可控、可用、可靠 [Si Xiao: Create an ethical "Ark" to make artificial intelligence knowable, controllable, usable and reliable]," Tencent Research Institute, Dec. 6, 2018, https://mp.weixin.qq.com/s/_CbBsrjrTbRkKjUNdmhuqQ; and "Li Yanhong Unveiled After 'Baidu Lost the Land,': "Simple Search" Without Advertisement, Mass Production of Unmanned Vehicles in July," *China IT News*, May 26, 2018, http://www.fonow.com/view/208592.html; "Beijing AI Principles," Beijing Academy of Artificial Intelligence, May 25, 2019, (archived by the Internet Archive on Nov. 4, 2020), https://web.archive.org/web/20201104201512/https://www.baai.ac.cn/news/beijing-ai-principles-en.html; and Graham Webster, "Translation: Chinese AI Alliance Drafts Self-Discipline 'Joint Pledge,'" New America, June 17, 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/.

40 James Vincent, "The Problem with AI Ethics," *The Verge*, April 3, 2019, https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech.

developed can be a powerful tool for guiding the future employment of AI technology. Because AI is so ubiquitous, many arms control measures will need to become widespread over time in order to be effective, as arms control is today for chemical and biological weapons. Efforts to engage in dialogue can start small. Were the United States and China — the world's major military and AI powers — to take the lead, it would be a powerful way to shape other countries' expectations about military AI.[41]

Shaping Norms

The United States has been proactively engaged in shaping norms about AI use. The U.S. government has released a steady stream of documents on military AI in recent years, including the Department of Defense's AI ethical principles, the Defense Innovation Unit's *Responsible AI Guidelines*, the Department of Defense's responsible AI strategy, an update to the Defense Department's policy on autonomy in weapons, a statement on human control over nuclear weapons in the 2022 *Nuclear Posture Review*, and a State Department political declaration on military AI use.[42] These unilateral policy statements will not constrain how other states develop AI, but they can help shape state views on how militaries might use AI. As a next step, U.S. policymakers should work with other states to adopt these principles to help shape emerging norms about AI use.

Building Cooperation

By going after "low-hanging fruit" — relatively unobjectionable arms control or confidence-building measures — states could help to build patterns of cooperation in order to manage AI risks. One area that is especially ripe for international collaboration is an "international autonomous incidents agreement" for uncrewed, autonomous vessels, drawing inspiration from the 1972 U.S.-Soviet Incidents at Sea Agreement.[43] Many existing international agreements already regulate the behavior of crewed aircraft and vessels, including the Convention on the International Regulations for Preventing Collisions at Sea, the Code for Unplanned Encounters at Sea, and multiple bilateral agreements between the United States and China.[44] Updating existing agreements or crafting a new agreement to cover uncrewed and autonomous systems could be a valuable step in building state trust and cooperation to help manage AI risks.

Shaping AI Development

The most important step that U.S. and allied policymakers can take today to control how AI is used in future conflicts is to shape the development of AI technology itself to make it more controllable. Computing hardware is an especially valuable point of control because of the trends in compute-intensive AI and the ability to physically limit access to chips. Policy decisions that are made today could make compute more or less governable in the future. The U.S. government has waded into industrial

41 Paul Scharre, "US and China Can Show World Leadership by Safeguarding Military AI," *South China Morning Post*, March 2, 2023, https://www.scmp.com/comment/opinion/article/3211478/us-and-china-can-show-world-leadership-safeguarding-military-ai.

42 "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. State Department, Feb. 16, 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/; *DOD Directive 3000.09: Autonomy in Weapons Systems*, Department of Defense, Jan. 25, 2023, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf; *Responsible Artificial Intelligence Strategy and Implementation Pathway*, Department of Defense, June 2022, https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf; *Responsible AI Guidelines in Practice*, Defense Innovation Unit, November 2021, https://assets.ctfassets.net/3nanhbfkr0pc/acoo1Fj5uungnGNPJ3QWy/3a1dafd64f22efcf8f27380aafae9789/2021_RAI_Report-v3.pdf; Kathleen Hicks, *Implementing Responsible Artificial Intelligence in the Department of Defense*, U.S. Defense Department, Memorandum, May 26, 2021, https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF; *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, Feb. 12, 2019, Department of Defense, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF; *DoD AI Ethics Principles*, Department of Defense, Feb. 24, 2020, https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/; *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, Defense Innovation Board, accessed April 25, 2023, https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF; and *2022 Nuclear Posture Review*, Department of Defense, October 2022, https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF, 13.

43 "Agreement Between the Government of The United States of America and the Government of The Union of Soviet Socialist Republics on the Prevention of Incidents On and Over the High Seas," U.S. Department of State, May 25, 1972, https://2009-2017.state.gov/t/isn/4791.htm.

44 "Convention on the International Regulations for Preventing Collisions at Sea, 1972," International Maritime Organization, Oct. 20, 1972, https://www.imo.org/en/About/Conventions/Pages/COLREG.aspx; "Code for Unplanned Encounters and Sea: Version 1.0," Western Pacific Naval Symposium, April 22, 2014, https://news.usni.org/2014/06/17/document-conduct-unplanned-encounters-sea; "Memorandum of Understanding Between the Department of Defense of the United States of American and the Ministry of National Defense of the People's Republic of China Regarding the Rules of Behavior for Safety of Air and Maritime Encounters," Department of Defense, 2014, https://dod.defense.gov/Portals/1/Documents/pubs/141112_MemorandumOfUnderstandingRegardingRules.pdf; and "Supplement to the Memorandum of Understanding on the Rules of Behavior for Safety of Air and Maritime Encounters Between the Department of Defense of the United States of America and the Ministry of National Defense of the People's Republic of China," Department of Defense, 2015, https://dod.defense.gov/Portals/1/Documents/pubs/US-CHINA_AIR_ENCOUNTERS_ANNEX_SEP_2015.pdf.

policy for semiconductors through both government subsidies and export controls, but often without a clear sense of what the policy goals are. Subsidies and export controls are important tools in laying the foundations for compute governance, but they are incomplete and could even be harmful if improperly executed. The U.S. government needs a comprehensive strategy to ensure continued control over access to large-scale computing resources. There are a number of key elements necessary for a successful compute governance strategy.

First, a comprehensive strategy ought to constrain China's production of advanced semiconductors. The October 2022 U.S. export controls on sending semiconductor manufacturing equipment to China dealt a heavy blow to China's domestic semiconductor manufacturing industry. But it will only succeed if Japan and the Netherlands join America in this effort. Collectively, Japan, the Netherlands, and the United States control 90 percent of the global market for semiconductor manufacturing equipment.[45] In early 2023, news reports indicated that Japan and the Netherlands adopted similar controls to the United States, although much will depend on which specific technologies these controls target.[46]

Second, the U.S. strategy should ensure that U.S. companies continue to dominate key chokepoints in the global supply chain. In its semiconductor industrial policy, the U.S. government should prioritize building a domestic ecosystem for leading-edge manufacturing in order to ensure that U.S. companies remain dominant in these important chokepoints for next-generation semiconductor manufacturing technology. Keeping U.S. companies dominant will ensure that the U.S. government retains the ability to control access to compute in the future.

Third, it is important to improve compute tracking to prevent the diversion of controlled chips to banned actors. U.S. export controls on AI chips will only be as effective as their enforcement. The U.S. government should improve its tools and resources for tracking and monitoring controlled chips to prevent banned actors from accumulating large amounts of compute.

Fourth, the U.S. strategy needs to keep Chinese firms dependent on compute resources that use U.S. technology. The goal of U.S. controls should not be to deny Chinese firms access to any U.S. technology but rather to keep them dependent on U.S. technology so that the U.S. government can control their access to large amounts of compute.[47] Policymakers should be mindful of potential downsides to export controls, particularly restrictions on AI chips themselves.[48] Export controls could accelerate the development of supply chains that do not rely on U.S. technology as states that are cut off from external sources redouble their efforts to grow their national capacity. Policymakers should advance policies that help retain centralized control over compute resources and thus the ability to restrict these resources in the future and not inadvertently accelerate their diffusion.

Fifth, Washington needs to enact export controls and cyber security requirements for trained models. Export controls on high-end chips will not be effective in constraining access to high-end AI capabilities if trained models are leaked, stolen, or released. Export controls on high-end chips must be supplemented with export controls on certain kinds of trained models in order for compute governance to be effective. Trained models could have security-relevant applications by design or as an emergent property of large, dual-use models. Export controls may be required for models trained for certain applications, such as for cyber security or generating new chemical compounds, because of their potential for misuse, or for any models above a certain compute threshold, because of their dual-use nature.[49] Similarly, U.S. companies conducting large-scale training runs should be required to comply with rigorous cyber security safeguards to ensure that their models are not stolen by malicious actors.

Finally, a comprehensive strategy needs to regulate large-scale compute use. In order to ensure that controls on advanced AI chips and trained models are effective, the U.S. government will need to enact a domestic regulatory regime to control the use of large amounts of compute. Otherwise, prohibited actors could simply access large-scale compute to train models through cloud providers. Regulations on compute use should cover large-scale training runs and the use of AI cloud computing centers.

45 Khan, Mann, and Peterson, *The Semiconductor Supply Chain*, 26.

46 Jenny Leonard and Cagan Koc, "Biden Nears Win as Japan, Dutch Back China Chip Controls," *Bloomberg*, Jan. 26, 2023, https://www.bloomberg.com/news/articles/2023-01-27/japan-netherlands-to-join-us-in-chip-export-controls-on-china#xj4y7vzkg.

47 Paul Scharre, "Decoupling Wastes U.S. Leverage on China," *Foreign Policy*, Jan. 13, 2023, https://foreignpolicy.com/2023/01/13/china-decoupling-chips-america/.

48 Sarah Bauerle Danzman and Emily Kilcrease, "The Illusion of Controls," *Foreign Affairs,* Dec. 30, 2022, https://www.foreignaffairs.com/united-states/illusion-controls.

49 For example, see Fabio Urbina et al., "Dual-Use of Artificial-Intelligence-Powered Drug Discovery," *Nature Machine Intelligence*, no. 4 (March 2022): 189–91, https://www.nature.com/articles/s42256-022-00465-9.epdf.

Regulations should, at a minimum, require reporting to U.S. regulators about large-scale training runs and cyber security standards for training runs above a certain compute threshold. They should also require cloud providers to ensure that their services are not used by prohibited actors.[50]

AI technology will continue to evolve rapidly, and those working on arms control initiatives must remain nimble and willing to adjust their focus to different aspects of AI technology or different AI-enabled military capabilities if the need arises. Keeping metrics for tracking AI progress and proliferation will be helpful in assessing possibilities for arms control as well as potential future challenges.[51]

At present, it is unclear how militaries will adopt AI, how the technology might affect warfare, and what, if any, forms of arms control states may perceive as desirable and feasible. There are actions that policymakers can take today, however, to lay the groundwork for potential arms control measures in the future, including not only shaping the technology's evolution but also the political climate. Actions taken today, even small ones, could yield large results down the road. States should seize these opportunities, when possible, to reduce the risks of military use of artificial intelligence.

Megan Lamberth *is a former associate fellow with the Technology and National Security Program at the Center for a New American Security.*

Paul Scharre *is the vice president and director of studies at the Center for a New American Security and is the author of* Four Battlegrounds: Power in the Age of Artificial Intelligence.

This article was made possible, in part, by the generous support of Open Philanthropy. This article is adapted from the authors' Center for a New American Security report, Artificial Intelligence and Arms Control.[52]

50 See also Jason Matheny, "Challenges and Opportunities for the Department of Defense: Testimony Presented to the U.S. Senate Committee on Armed Services, Subcommittee on Cybersecurity, on April 19, 2023," RAND Corporation, April 19, 2023, https://www.armed-services.senate.gov/imo/media/doc/CTA2723-1_v2.pdf.

51 "The AI Index," https://aiindex.stanford.edu/; and Jess Whittlestone and Jack Clark, "Why and How Governments Should Monitor AI Development," arXiv, Aug. 28, 2021, https://doi.org/10.48550/arXiv.2108.12427.

52 Paul Scharre and Megan Lamberth, "Artificial Intelligence and Arms Control," Center for a New American Security, Oct. 12, 2022, https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control.