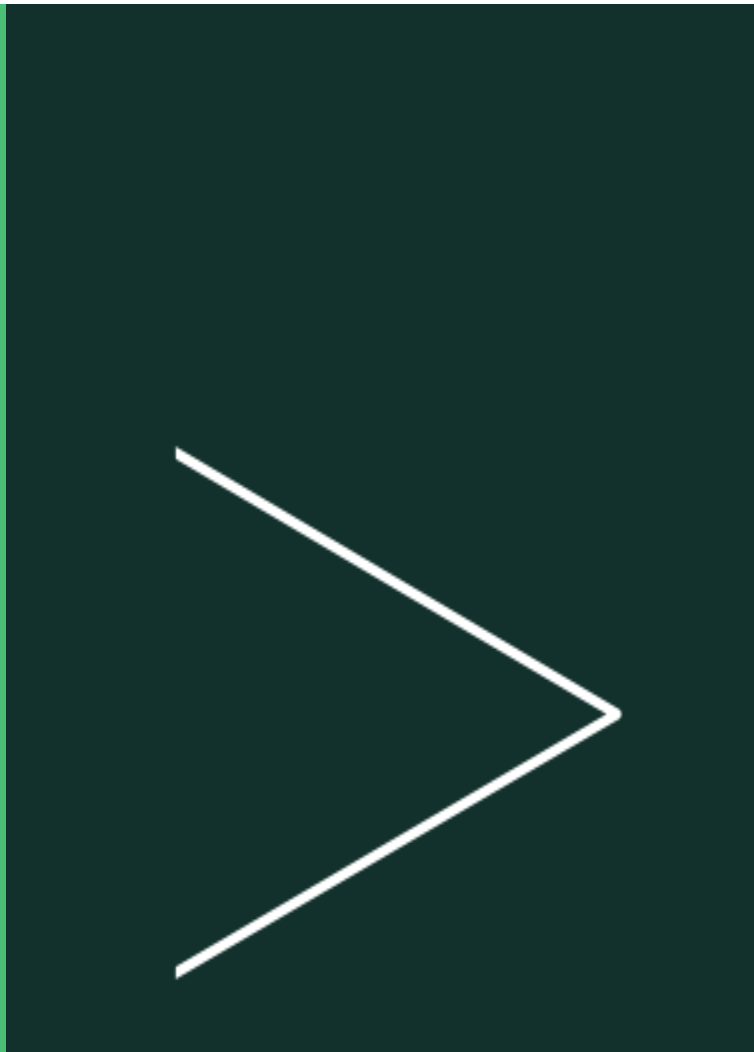
Founders Pledge

# Autonomous Weapon Systems & Military AI

**CAUSE AREA INVESTIGATION**

CHRISTIAN RUHL

May 2022



# Contents


**What is the Problem?** **4**

Military Benefits of Autonomous Weapons 5

Increasing Autonomy and Military AI Investments 5

**Pathways to Risk** **6**

Pathway I: "Slaughterbot" Scenarios 8

Pathway II: WMD Delivery Systems 11

Pathway III: War at Machine Speed 14

Counter-Argument and Response 15

Pathway IV: Automation Bias 17

Pathway V: Direct Strategic Stability Effects 19

Pathway VI: Systems Complexity 21

Pathway VII: Military AI Competition 23

**Unpacking the Risk** **25**

Nuclear Close Calls 25

Setting Precedents for AI Regulation 27

**The Funding Landscape: Neglectedness and Diversification** **28**

**Evaluating Interventions** **30**

The Problems with a "Killer Robot Ban" 30

Focus on Key Stakeholders, Not Multilateralism 31

**Funding Research on Strategic Risks of AWS** **32**

Field-Building 34

Actionable CBM Options 37

Track II Dialogues 38

International Autonomous Incidents Agreement 39

**Conclusion** **41**

**Disclosures** **41**

# Summary

The use and proliferation of autonomous weapon systems appears likely in the near future, but the risks of AI-enabled warfare are under-studied and under-funded. Autonomous weapons and military applications of AI more broadly (such as early-warning and decision-support systems) have the potential to increase the risk factors for a variety of issues, including great power war, nuclear stability, and AI safety. Several of these issues are potential pathways towards existential and global catastrophic risks. Autonomy in weapon systems therefore affects both the long-term future of the world and the lives of billions of people today.

This report intends to advise philanthropic donors who wish to reduce the risk from autonomous weapon systems and from the military applications of AI more broadly. We argue that much of the problem arises from strategic risks that affect the likelihood of great power conflict, nuclear war, and risks from artificial general intelligence. These risks include the increased speed of decision-making in a world with autonomous weapons, automation bias, increased complexity leading to a higher risk of accidents and escalation, and the possibility of AI-related military competition and its implications for long-term AI safety.

Although “killer robots” feature in the popular imagination, and some prominent organizations have taken up and promoted the cause, autonomous weapons remain a neglected issue for three reasons. First, the largest organizations focus mostly on humanitarian issues, leaving strategic threats relatively neglected. Second, those who do study risks beyond “slaughterbots” (like automation bias, strategic stability, etc.) are few and receive even less funding; there is a talent shortage and room for funding. Third, the most widely-advocated solution — formal treaty-based arms control or a “killer robot ban” — is not the most tractable solution. Thus, philanthropists have an opportunity to have an outsized impact in this space and reduce the long-term risks to humanity’s survival and flourishing.

In addition to outlining the risks from autonomous weapon systems and the military applications of AI, we also evaluate potential interventions to mitigate these risks. We argue that philanthropists can have an outsized impact in this space by following two guiding principles to choose their interventions:

1. **Focus on strategic risks**, including the effects of autonomous systems on nuclear stability.
2. **Focus on key actors, rather than multilateral inclusiveness**, and prioritize those states most likely to develop and use autonomous systems, possibly starting with bilateral dialogues.

Using these guiding principles, we argue that effective philanthropists should focus not on a legally binding ban of autonomous weapons, but on researching strategic risks and on working on a set of so-called confidence-building measures or CBMs, which have a track record of regulating militarily-useful technologies. Funding work related to this research and CBMs presents one of the best ways to have an impact on this important and neglected problem.

***Acknowledgments**: The author would like to thank Anthony Aguirre, Stephen Clare, Sjir Hoeijmakers, Emilia Javorsky, Matt Lerner, Carl Robichaud, and Shaan Shaikh for their helpful comments and advice on earlier drafts, and to thank Paul Scharre and Michael Horowitz, whose research and insights on AI-enabled military systems are foundational to much of this report.*

## What is the Problem?

An autonomous weapon system (AWS) can be defined as "[A] weapon system that, once activated, is intended to select and engage targets where a human has not decided those specific targets are to be engaged."[1] Although such systems sound like science fiction — a perception exacerbated by some media portrayals — all component parts to create autonomous weapons exist today, and existing weapon systems possess high degrees of automation.[2] These systems use artificial intelligence (AI) or interact with AI-enabled systems (like early warning, decision support, command and control, and intelligence-analysis systems), which is why this report will refer to the terms "autonomous weapon system," "military AI," and "AI-enabled weapon system," to describe facets of the same issue of delegating military decisions to machines. Note that this report largely focuses on risks from existing and near-future narrow AI systems, not superintelligent or general AI, though we believe the problems are interconnected, as discussed in "Setting Precedents for AI Regulation," below.

As Dr. Paul Scharre, one of the key scholars of autonomy in weapon systems, has explained, autonomous weapons can take many different forms, including but not limited to drone swarms, as demonstrated by swarm-on-swarm autonomous dogfights conducted at the Naval Postgraduate School; autonomous ships, like DARPA's Sea Hunter; or any AI-enabled weapon or weapons system (i.e. not just "slaughterbot" drone swarms).[3] This report embraces the diversity of systems that might be considered part of the autonomy spectrum, and is intended as a guide to philanthropists rather than an exercise in classifying weapons or defining "intelligence" or "autonomy." Not all of these threats are new; as we explain below, and as other scholars have pointed out, some concerns about "slaughterbots" differ little from long-standing concerns about the proliferation of cheap arms to rogue actors in the international system. Other threats discussed in this report appear to alter the character of war in surprising and risky ways, and potentially undermine nuclear stability. It is these threats that we believe pose the most important risks from autonomous weapons and near term narrow AI in military systems.

---

[1] Paul Scharre and Michael Horowitz, *An Introduction to Autonomy in Weapon Systems*, Project on Ethical Autonomy (Washington, DC: Center for a New American Security, 2015), 5, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems. Note that there is an ongoing definitional debate among scholars and activist groups, as outlined ibid., 3.
[2] For instance, as Horowitz has explained, a non-state armed group could combine a self-driving vehicle with a machine gun, commercially-available AI, and heat-tracking sensor, and then "write software code such that the machine gun will fire at anything that has the heat signature of a human being. While such a system would be indiscriminate and violate the Law of War in nearly all use cases, it is buildable today." Michael C. Horowitz, "When Speed Kills: Lethal Autonomous weapon systems, Deterrence, and Stability," *Journal of Strategic Studies* 42, no. 6 (2019): 778, https://doi.org/10.1080/01402390.2019.1621174.
[3] Scharre, *Army of None,* 11-13, 79.

## Military Benefits of Autonomous Weapons

We think there are major risks to autonomous weapon systems, which we explain in the next section. However, we first turn briefly to the apparent benefits (at least to some militaries) of such weapon systems. We believe a dispassionate assessment of costs and benefits is crucial for a rational understanding of the risks arising from AI-enabled warfare. After all, states, too, have strong incentives to avoid unnecessary risks. Thus, if autonomous weapons were simply risky systems with few upsides for military operations, then they would be unlikely to be adopted widely in the first place, and the scope of the problem would be small.[4]

What are the benefits that some states and militaries may see in autonomous weapons? They include:

1. **Speed, Surprise, and Survivability**: E.g. "An autonomous plane might be more adept at identifying and avoiding air defense threats, for example, or better at predicting and defeating adversaries in an air-to-air dogfight, making it more able to complete its mission."[5]
2. **Increased accuracy and decreased collateral damage**: With greater computing power, better access to information, and more speed than humans, autonomous weapons could more accurately and quickly calculate and pursue actions that increase accuracy and minimize collateral damage.[6]
3. **Machine-like adherence to rules and law**: Future autonomous weapon systems could be programmed to follow the laws of war and other rules of engagement more accurately than humans.[7] Autonomous weapon systems lack the kinds of human emotional shortfallings that lead to outbursts of violence or even massacres.[8]
4. **Decreasing civilian harm**: Future autonomous systems may outperform humans in distinguishing combatants from civilians, thereby decreasing civilian casualties in war.[9]
5. **Removing humans from the battlefield**: Future wars fought with uninhabited machines would spare some human lives. For example, mine-defusing robots are already saving human lives by delegating dangerous jobs to non-human agents.[10]

## Increasing Autonomy and Military AI Investments

States with advanced military and robotics capabilities — including the United States and China — are thus investing heavily in the military applications of AI, and government strategic planning documents, like the U.S. DoD's Unmanned Systems Roadmap, support a drive for greater autonomy.

---

[4] Thanks to Professor Michael Horowitz for pointing this out.
[5] Horowitz, "When Speed Kills," 769.
[6] Wendell Wallach and Colin Allen, "Framing Robot Arms Control," *Ethics and Information Technology* 15, no. 2 (June 2013): 128, https://doi.org/10.1007/s10676-012-9303-0.
[7] For an overview and challenges, see ibid.
[8] "… and unlike human soldiers, machines never get angry or seek revenge," Scharre, *Army of None*, 6.
[9] "It isn't hard to imagine future weapons that could outperform humans in discriminating between a person holding a rifle and one holding a rake." Scharre, *Army of None*, 6.
[10] Scharre, *Army of None*, 14.

[11] To quantify this drive, according to the Center for Security and Emerging Technology, the FY 2021 U.S. Defense Budget Request included $1.7 *Billion* of investments in "autonomy to enhance 'speed of maneuver and lethality in contested environments' and the development of 'human/machine teaming,'" in addition to other investments in the military application of AI.[12] So far, the U.S. does not have operational lethal autonomous weapons in its inventory, but they are actively developing AI-enabled systems and appear open to deploying autonomous weapons if adversaries do.[13]

In short, the advent of truly autonomous weapon systems appears likely in the near future. Crowdsourced probabilistic forecasts support this prediction. Metaculus predicted the probability of assassination by autonomous weapon by 2025 as 12%.[14] Loitering munitions, which have some degrees of autonomy, have already been used on the battlefield (e.g. in Libya, according to a UN panel), although these weapons are not new and do not exhibit high degrees of autonomy.[15] To better understand the nature of this risk, the next section dives deep into several potential pathways to risk, based on a review of the literature on autonomous weapons.

## Pathways to Risk

Autonomous weapon systems act as a "threat multiplier," activating pathways for a variety of risks as outlined in figure 1, below. These risks — like great power conflict and thermonuclear war — could lead to the deaths of millions or even billions of people, and in the worst cases, to the extinction of humanity or the unrecoverable loss of our potential.[16] This section dives into greater detail for each of the potential risks that autonomous weapons could pose to international security and the future of humanity. These pathways are not exhaustive of the full spectrum of risk, but help to illustrate the various ways that scholars and experts on this issue currently believe autonomous weapons could pose major risks to humanity.

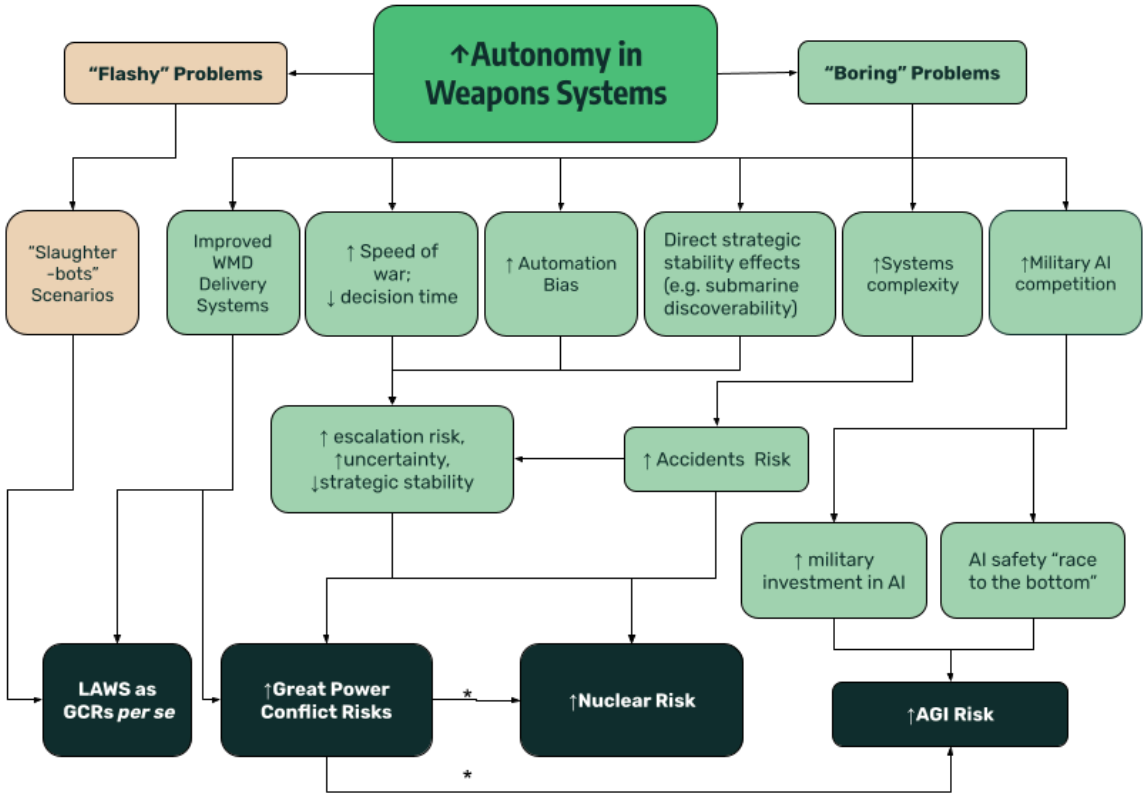
The seven pathways described below include:

---

[11] Jürgen Altmann and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability," *Survival: Global Politics and Strategy* 59, no. 5 (2017): 122, https://doi.org/10.1080/00396338.2017.1375263.
[12] CSET, "U.S. Military Investments in Autonomy and AI: A Strategic Assessment," *Center for Security and Emerging Technology*, https://cset.georgetown.edu/publication/u-s-military-investments-in-autonomy-and-ai-a-strategic-assessment/.
[13] "Although the United States does not currently have LAWS in its inventory, some senior military and defense leaders have stated that the United States may be compelled to develop LAWS in the future if potential U.S. adversaries choose to do so." Congressional Research Service, "Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems," *Federation of the American Scientists*, last modified December 2020, accessed November 12, 2021, https://sgp.fas.org/crs/natsec/IF11150.pdf.
[14] Until recently, CSET Foretell also hosted a question on autonomous weapons, but has now transitioned to INFER-pub, and the question is no longer active.
[15] Joe Hernandez, "A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says," *NPR,* https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d
[16] For more on this argument, see John Halstead, Safeguarding the Future Cause Area Report (Founders Pledge, 2020), https://founderspledge.com/stories/existential-risk-executive-summary or Toby Ord's *The Precipice.*

- Pathway I: “Slaughterbot” Scenarios
- Pathway II: WMD Delivery Systems
- Pathway III: War at Machine Speed
- Pathway IV: Automation Bias
- Pathway V: Direct Strategic Stability Effects
- Pathway VI: Systems Complexity-Related Risks
- Pathway VII: Military AI Competition

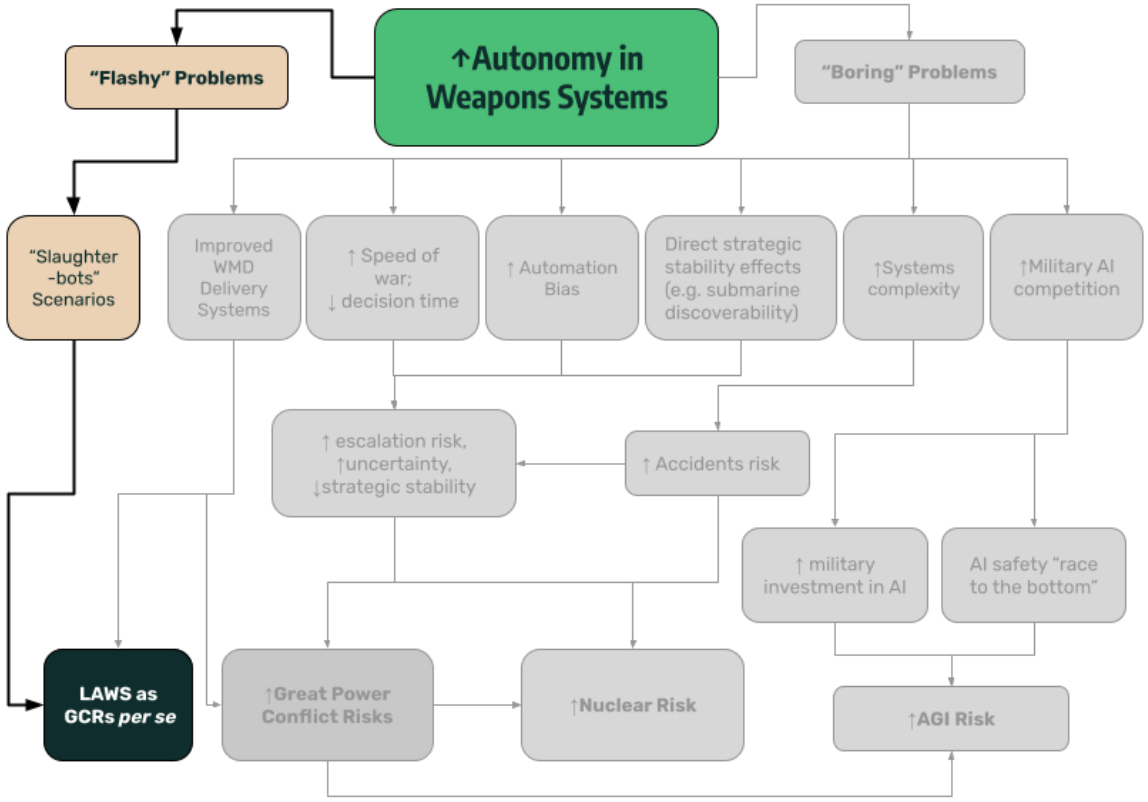
Figure 1: The Autonomous Risk Landscape




*Orange = “flashy” problems; light green = “boring” (but neglected) problems; dark green = GCRs and existential threats. Source: author’s diagram based on literature reviewed below. *= see Great Power War report.*

## Pathway I: “Slaughterbot” Scenarios

Figure 1.1. “Flashy” Risk Path




The first potential risk is that of so-called “slaughterbots,” or armed drones with target-seeking and swarming behavior. Some scholars have argued that drone swarms could be classified as weapons of mass destruction (WMD).[17] Anthony Aguirre, an associate professor of physics at UCSC who has worked on autonomous weapons with the Future of Life Institute, for example, has argued that slaughterbots are cost-effective killing mechanisms compared to other WMD like nuclear weapons, and are potentially more likely to be used, given their precision and the advantages outlined above. Aguirre calculates the “cost-effectiveness” of military systems:

[17] Zachary Kallenborn, *Are Drone Swarms Weapons of Mass Destruction?,* Future Warfare Series 60 (United States Airforce Center for Strategic Deterrence Studies, 2020), https://media.defense.gov/2020/Jun/29/2002331131/-1/-1/0/60DRONESWARMS-MONOGRAPH.PDF. and Aguire, "Why those who care about catastrophic and existential risk should care about autonomous weapons," Effective Altruism Forum (blog), entry posted November 11, 2020, https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2#fnref-PvLbqbJx5AokdjcWu-13.

$$(manufacturing\ cost\ +\ deployment\ cost) \div fatality\ rate\ =\ cost\ per\ fatality$$

With some assumptions ($100/unit cost for mass manufacturing, the same cost for deployment, and a 50% fatality rate), this leads Aguirre to suggest:

$$(\$100\ +\ \$100) \div 0.5\ =\ \$400\ per\ fatality$$[18]

For 100,000 fatalities (comparable to a nuclear strike on a medium sized city), the cost would therefore be $40 Million and expected to drop in the future, Aguirre argues. As a very rough (and possibly invalid) comparison, North Korea is believed to produce enough fissile materials for about 6-7 nuclear weapons a year,[19] and South Korean government experts estimate it spends roughly between $1.1 billion and $3.2 billion on its nuclear weapons program per year as of 2017,[20] leading to a rough cost of between $157 million and $533 million per new warhead (neglecting the cost of maintaining and deploying current warheads and missiles).

In general, a more complete estimate of a weapon's cost would also need to incorporate initial costs of research, development, and testing. In the case of AI, however, many of these costs have been outsourced to the private sector, and we therefore believe that Aguirre's estimates are still valuable for a thought experiment. It should be noted that this concept of cost-effectiveness does not necessarily reflect military thinking — "cost per fatality" is usually not the relevant measure, as much as "ability to take and hold territory" (for infantry), "ability to ensure access to waterways" (for warships), etc.[21] It may apply better to the calculus of non-state actors like terrorist groups.

This calculation does, however, illustrate that components for basic AI-enabled systems are cheap and commercially available, and therefore suggest "slaughterbot" proliferation risk.[22] Buying relatively cheap drones and code (e.g. swarming algorithms and facial recognition software) may enable some actors to compensate for military weakness, destabilizing the international system. It is unclear to us whether this destabilizing threat is unique to AI-enabled weapons, or is simply an issue of drone proliferation and the cost of hardware generally. More research is needed to better understand how this kind of proliferation may affect the global order and the risk of war.

In spite of the apparent cost-effectiveness of slaughterbots as weapons of mass destruction, however, there exist defensive measures that may help defend against this risk. For example, as

[18] Aguirre, "Why those who care about catastrophic and existential risk should care about autonomous weapons."
[19] Arms Control Association, "Nuclear Weapons: Who Has What at A Glance," *Arms Control Association,* January 2022. https://www.armscontrol.org/factsheets/Nuclearweaponswhohaswhat
[20] James Pearson and Ju-Min Park, "North Korea overcomes poverty, sanctions with cut-price nukes," *Reuters,* 11 January 2016, https://www.reuters.com/article/us-northkorea-nuclear-money/north-korea-overcomes-poverty-sanctions-with-cut-price-nukes-idUSKCN0UP1G820160111.
[21] Thanks to Shaan Shaikh for highlighting the importance of this in a round of external reviews of this report.
[22] Thanks to Emilia Javorsky and Anthony Aguirre from the Future of Life Institute for their comments on proliferation risk in a round of external reviews of this report.

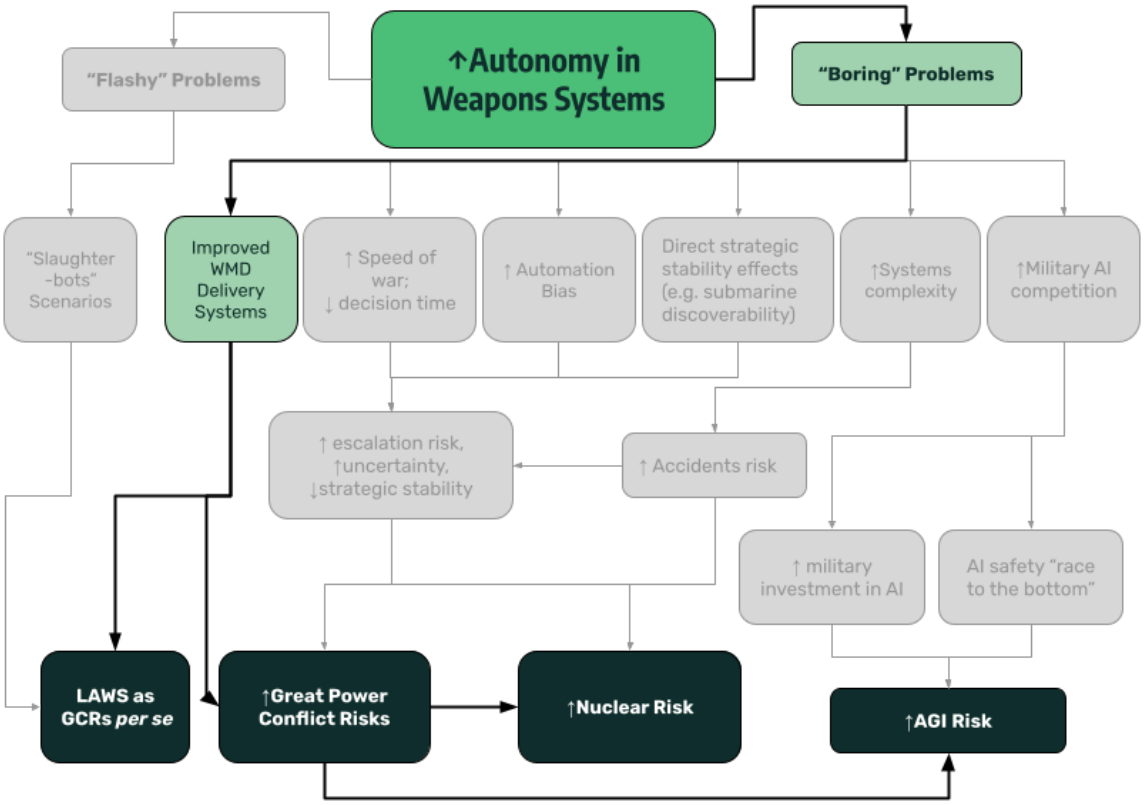


Paul Scharre has explained, swarms are susceptible to simple communications jamming to "collapse the swarm," and perhaps even to take over cyber operations that turn a swarm on its originator.[23] Other defensive measures like lasers and rail guns, and even low-tech solutions like swarm-catching nets are being discussed and developed.[24] Thus, we categorize slaughterbots as a "flashy" problem — although they receive much more media attention (like coverage of the famous 2017 "Slaughterbots" video) and funding (discussed below) than the other risks of this report, possible defenses appear more likely, and the other issues discussed below are more neglected.

[23] Paul Scharre, "Counter-Swarm: A Guide to Defeating Robotic Swarms," *War on the Rocks*, March 31, 2015, accessed November 12, 2021, https://warontherocks.com/2015/03/counter-swarm-a-guide-to-defeating-robotic-swarms/.
[24] Ibid.

## Pathway II: WMD Delivery Systems

Figure 1.2. Risks from AWS as Delivery Systems




One pathway towards existential risk is that autonomous weapons could be developed as delivery systems for other weapons of mass destruction. This development that may be especially worrying regarding the capabilities of rogue states and nonstate actors like terrorist groups, and genocidal regimes using facial-recognition technology. For example, as one scholar of autonomous drones put it:

> Autonomous vehicles are a great[25] way to deliver chemical, radiological, and biological weapons. An autonomous vehicle cannot get sick with anthrax, nor choke on chlorine. Drones can more directly target enemies, while adjusting trajectories based on local wind and humidity conditions. Plus, small drones can take to the air, fly indoors, and work together to carry out attacks. Operatives from the Islamic State in Iraq and Syria were reportedly quite interested in using drones to carry out radiological and potentially chemical attacks. North

[25] We would say "potentially effective."



Korea also has an arsenal of chemical, biological, and nuclear weapons and a thousand-drone fleet.[26]

It should be noted that these points are not necessarily unique to AI-enabled systems and may be applicable to remotely-piloted systems more generally (e.g. non-autonomous drone swarms delivering WMD, controlled from another country). Moreover, the examples of ISIS and North Korea do not involve significant autonomy or AI-enabled capabilities as far as we know.[27] Researchers at the Stockholm International Peace Research Institute have highlighted that AI-enabled weapon systems could have destabilizing effects when applied to nuclear delivery, writing, "Advances in machine learning and autonomy open up the possibility of using new types of platform for nuclear delivery, notably UAVs, UUVs and hypersonic glide vehicles" and that such systems "are more prone to loss of human control due to malfunction, hacking or spoofing, which could lead to accidental nuclear weapon use."[28]

Automation is not new in nuclear delivery either. For example, intercontinental ballistic missiles (ICBMs), once launched, navigate to and strike their targets without further human intervention, as scholars of AI-related nuclear risk have pointed out.[29] Moreover, the logic outlined here runs both ways again — if autonomous weapons-delivery platforms are very risky, states have incentives to avoid adopting them. This may explain why senior U.S. military officials have made statements like "I like the man in the loop," expressing a preference for human control of nuclear delivery, a preference that is also reflected in the National Security Commission on AI's *Final Report*.[30] Despite this logic, some states apparently do view the potential of AI-enabled delivery platforms as a benefit; Russia is investing in its "Poseidon" or "Status 6" system, an uninhabited nuclear-capable underwater vehicle that will reportedly have autonomous capabilities.[31] Part of the reason may be to evade defensive systems: "Status-6 might be able to use AI to evade enemy anti-submarine warfare (ASW) forces on the way to its target."[32]

If, as discussed above, autonomous systems are more prone to unpredictable behavior or "malfunction, hacking, or spoofing," then the risk of accidents and unintended escalation may

[26] Zachary Kallenborn, "A Partial Ban A Partial Ban on Autonomous Weapons Would Make Everyone Safer," *Foreign Policy,* https://foreignpolicy.com/2020/10/14/ai-drones-swarms-killer-robots-partial-ban-on-autonomous-weapons-would-make-everyone-safer/.
[27] Thanks to Shaan Shaikh for pointing this out.
[28] Vincent Boulanin, "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI,* 109, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf
[29] Michael Horowitz, Paul Scharre, and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence* (2019), 20, https://arxiv.org/pdf/1912.05291.pdf.
[30] Ibid., quoting General Robin Rand; NSCAI *Final Report,* 97, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.
[31] Vincent Boulanin, "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI,* 27, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf
[32] Michael Horowitz, Paul Scharre, and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence* (2019), 20, https://arxiv.org/pdf/1912.05291.pdf.

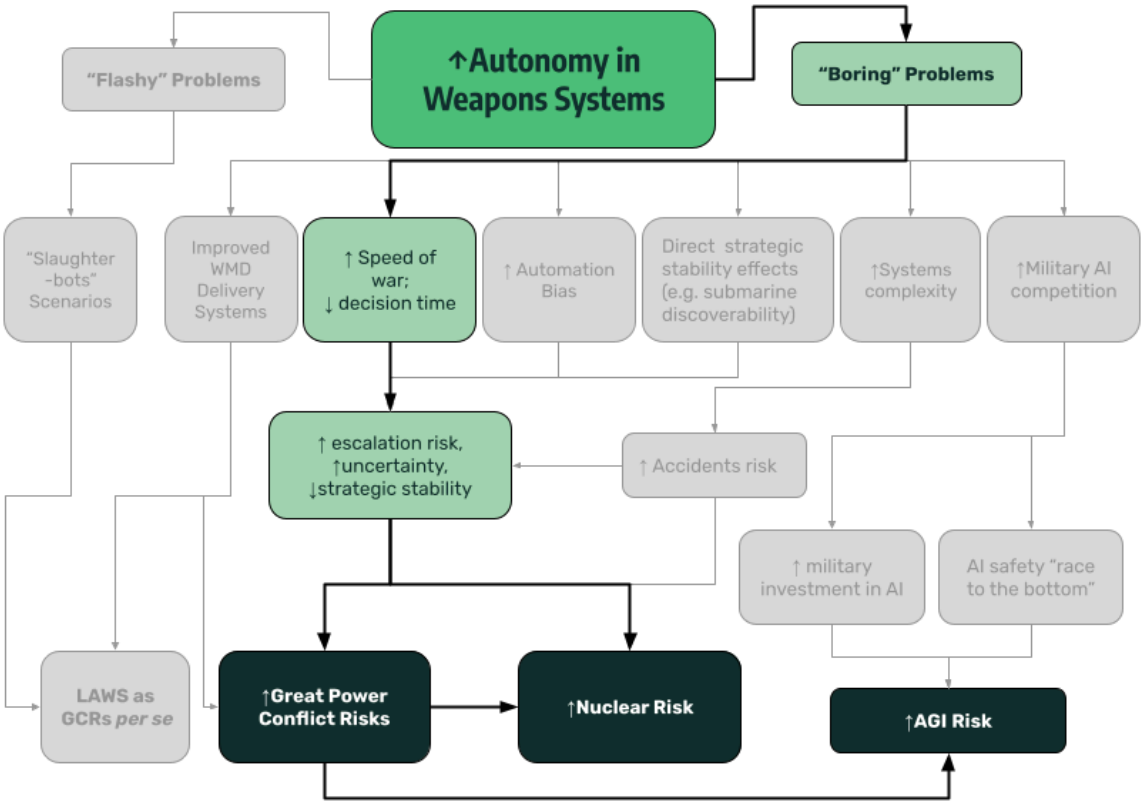
increase.[33] Moreover, unlike some similar risks below — like the section on strategic stability risks — improved WMD delivery systems may not only destabilize deterrence between nuclear-armed states, but may also amplify risks from non-state actors and rogue states. As explained above, drone swarms are potentially cheap; programmable drones are commercially available, and are thus a readily-available method that does not endanger a terrorist group's own operatives when delivering a radiological, biological, or chemical attack. Parts of this threat exist with or without AI[34] when even "dumb" drones can be used in a chemical or biological attack, while other parts — like using advanced facial-recognition software to target only certain groups — may be unique to more advanced autonomous systems. This may depend on the weapons being delivered (comparing, e.g. a highly contagious agent like weaponized plague with a more targeted weapon like anthrax that is not spread from person to person). More research is needed to fully understand the WMD-delivery risks, and we recommend research on strategic risks as a potential intervention, below.

[33] Vincent Boulanin, "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI,* 109, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf

[34] Thanks to Shaan Shaikh for emphasizing this point in comments on an earlier draft of this report.

## Pathway III: War at Machine Speed

Figure 1.3. Risks from Speed




A third risk from increased automation in weapon systems is that it increases the speed of war to "machine speed" because autonomous systems can process information and make decisions more quickly than humans.[35] This could decrease decision-making time, which in turn, might increase the risks of accidents and inadvertent escalation, both on the battlefield, and in strategic decisions. The advantages of speed are driving great power decision-making to the extent that military theorists and policy-makers from both the United States and China predict a new kind of war characterized by extreme speeds beyond human control. This is known as "hyperwar" in the West and "battlefield singularity" in China.[36] As political scientists Michael Horowitz and Paul Scharre point out, such dynamics could lead to an arms race for speed, even if countries would prefer to avoid fighting hyperwar: "The logic of a battlefield singularity, or hyperwar, is troubling precisely because competitive pressures could drive militaries to accelerate the tempo of operations and remove

[35] Paul Scharre, "A Million Mistakes a Second," *Foreign Policy,* https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/
[36] Michael C. Horowitz and Paul Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, CNAS, 5.

humans 'from the loop,' even if they would rather not, in order to keep pace with adversaries."[37] Senior policymakers like former Deputy Secretary of Defense Robert O. Work have echoed this concern.[38]

Consider, for example, the famous case of the Soviet Lieutenant Colonel Stanislav Petrov, who decided to ignore the early warning system telling him with high confidence that an ICBM strike was on its way to Russia. Petrov decided to wait and search for corroborating evidence rather than alert his superiors — human judgment and time allowed him to avert a nuclear war.[39] If Petrov's job was delegated to an autonomous system, or if autonomous systems were integrated in the entire Russian nuclear command-and-control system, there would have been no time to wait, and an all-out nuclear war could have started.[40] As the game theorist Thomas Schelling put it, "the premium on haste" is "the greatest source of danger that peace will explode into all out war."[41] Note that, as pointed out in Founders Pledge's recent Great Power Conflict report, an increased risk of great power conflict also affects nuclear risk and AGI risk.

A less well-known example from nuclear history also helps to illustrate the risks of AI-enabled speed in war-time decision-making. On November 9, 1979, a technician accidentally inserted a training tape with a simulation of a large-scale nuclear attack into a computer, and as a result, North American Aerospace Defense Command (NORAD) told Strategic Air Command (SAC) that it detected a large-scale attack.[42] "Within minutes, U.S. intercontinental ballistic missile (ICBM) crews were put on highest alert, nuclear bombers prepared for takeoff, and the National Emergency Airborne Command Post—the plane designed to allow the U.S. president to maintain control in case of an attack—took off."[43] After waiting for *six minutes* for satellite confirmation, and failing to receive it, officials decided the threat was not real.[44] In a world with widespread AI-enabled systems, where many processes are faster and more automated, decision-makers might feel they do not have the luxury of waiting for six minutes.

## Counter-Argument and Response

Some scholars have pointed out that, *if managed correctly,* increased speed could theoretically *lengthen* decision-making time.[45] If automation and AI-enabled early-warning systems can shorten the window needed to detect an incoming attack (before making a decision), and shorten the

[37] Ibid.
[38] Work: "If our competitors go to Terminators ... and it turns out the Terminators are able to make decisions faster, even if they're bad, how would we respond?" quoted ibid., 5-6.
[39] For Petrov as an example related to AWS risk, see Scharre, *Army of None,* 1-2, 207.
[40] Ibid, 207.
[41] Schelling, quoted ibid., 305.
[42] Future of Life Institute, "Accidental Nuclear War: A Timeline of Close Calls," Future of Life, https://futureoflife.org/background/nuclear-close-calls-a-timeline/.
[43] Ibid.
[44] Ibid.
[45] Michael Horowitz, Paul Scharre, and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence* (2019), 14, https://arxiv.org/pdf/1912.05291.pdf.



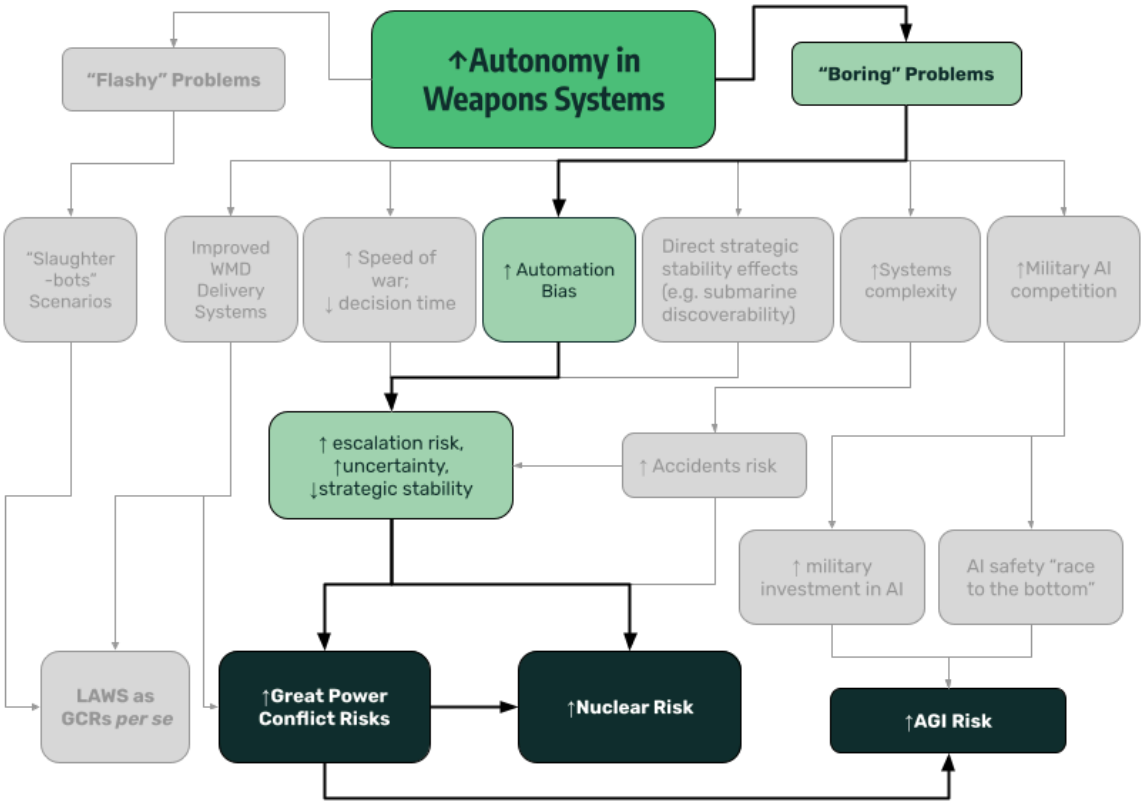
window needed to launch an attack (after making a decision), then decision-making time could theoretically be increased.

We believe, however, that this logic may not necessarily apply, because competitive launch pressures still favor an *overall* shortened decision-making window. As explained above, speed can be an advantage in war. In the minds of decision-makers, therefore, any second shaved off the overall launch process could be a competitive advantage over adversaries. This implies that increased speed in some parts of the launch decision-making process does not necessarily lead to increased deliberation time.

Lengthening the primary decision window would need to be a conscious choice to resist the competitive pressures favoring speed, and instead turn toward good judgment, better decision-making, and increased safety. Thus, we believe this overall counterargument should be only a small update against the risks of speed.

## Pathway IV: Automation Bias

Figure 1.4. AWS Risks from Automation Bias




A fourth risk — automation bias in war — arises from the cognitive biases that exist when humans interact with machines. Other scholars, including Horowitz and Scharre, have examined the implications of this bias in AI-enabled warfare in greater depth.[46] Put simply, in human-machine pairings, humans are likely to be biased towards the judgment of the machine, even when there is contradicting evidence. In other words, automation bias is defined as the tendency to "over-accept computer output 'as a heuristic replacement of vigilant information seeking and processing.'"[47] Automation bias has been recorded in a variety of areas, including medical decision-support systems, flight simulators, air traffic control, and even "making friendly enemy engagement decisions" in shooting-related tasks.[48] In a world with high autonomy in weapon systems and many

[46] Ibid., 4.
[47] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association* 19, no. 1 (January 2012): 121, https://doi.org/10.1136/amiajnl-2011-000089.
[48] David Lyell and Enrico Coiera, "Automation Bias and Verification Complexity: A Systematic Review," *Journal of the American Medical Informatics Association* 24, no. 2 (March 1, 2017): 423–31, https://doi.org/10.1093/jamia/ocw105.

human-machine teams, human rationality may be impaired significantly by such bias. A real life example from the Iraq war confirms the relevance of this issue:

> "Automation bias was operationally seen in the 2004 war in Iraq when the U.S. Army's Patriot missile system, operating in a management-by-exception mode,[49] engaged in fratricide, shooting down a British Tornado and an American F/A-18, killing three. [...] Given the laboratory evidence that given an unreliable system, humans are still likely to approve computer-generated recommendations [...], it is not surprising that under the added stress of combat, Patriot operators did not veto the computer's solution."[50]

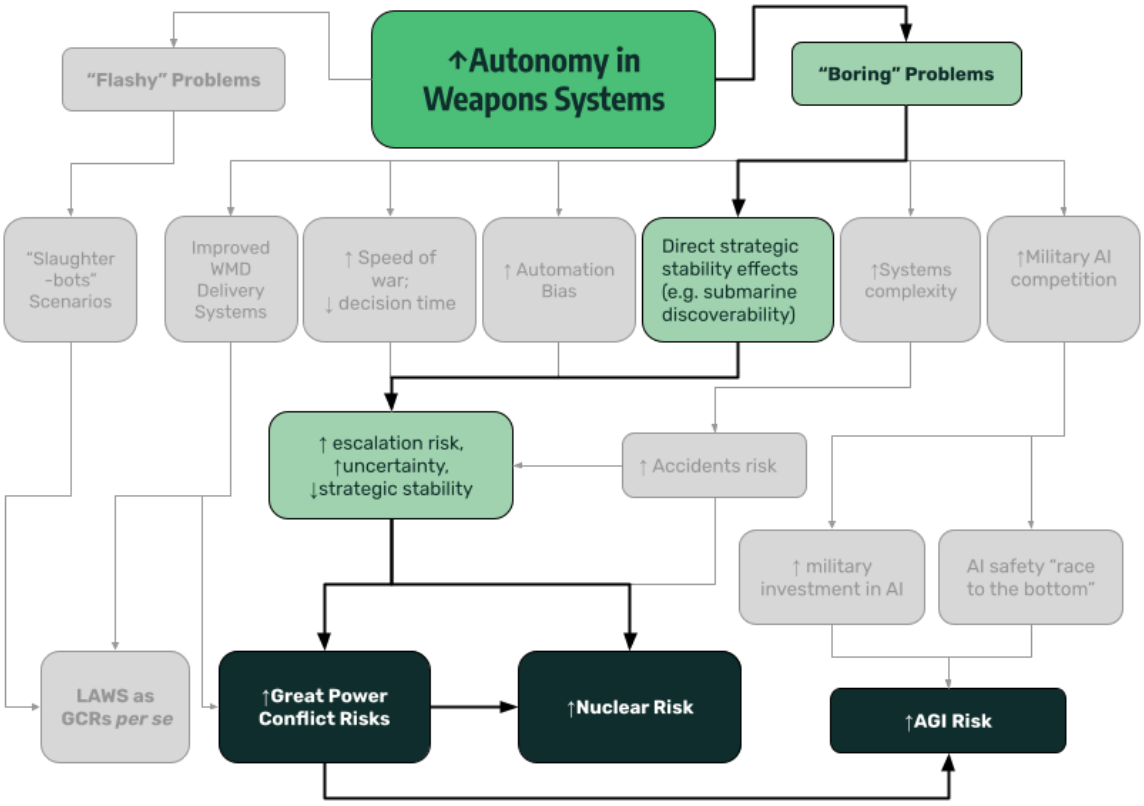
Similar dynamics leading to accidents and increased escalation risks can operate on the strategic level, too, and Petrov is again an illustrative example. In a world with more autonomous systems, humans might be biased against detecting false alarms and bugs in AI-enabled lethal systems, including "smarter" early warning systems informing a modern-day Petrov.[51] Unfortunately, building more sophisticated and less error-prone AI systems may not solve the problem of automation bias, and may even exacerbate it. In the words of one systematic review of automation bias, "high levels of system accuracy may inadvertently contribute to AB [automation bias]. This may be because accuracy engenders trust, and it has been shown that users who have greater trust in automation are less likely to detect automation failures."[52]

For these reasons, automation bias may turn out to be one of the most insidious "boring" risks of increased autonomy in weapon systems, and of automation more generally. Nonetheless, like the driverless car debate — where some arguments rest in part on the relative frequencies of human mistakes and machine mistakes — there is ultimately an empirical question, as human operators also make plenty of mistakes without AI:[53] Do human-machine pairs make fewer mistakes in critical situations than humans alone? The answers to such questions would have high value of information, which is in part why we emphasize research on strategic risks as a high-value intervention below.

[49] "Management by exception" in this case meant that the computer gave the human supervisor a set amount of time to oppose its proposed action, and executed the action if the human did not oppose.
[50] M.L. Cummings, S. Bruni, S. Mercier, and P.J. Mitchell, "Automation Architecture for Single Operator, Multiple UAV Command and Control," *The International C2 Journal 1,* no. 2 (2007): 7-8; also cited and discussed in Michael Horowitz, Paul Scharre, and Alexander Velez-Green, *A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence* (2019), 8.
[51] Michael Horowitz, Paul Scharre, and Alexander Velez-Green, *A Stable Nuclear Future?*, 4.
[52] David Lyell and Enrico Coiera, "Automation Bias and Verification Complexity: A Systematic Review," *Journal of the American Medical Informatics Association* 24, no. 2 (March 1, 2017): 423, https://doi.org/10.1093/jamia/ocw105.
[53] Thanks again to Shaan Shaikh for pointing this out.

## Pathway V: Direct Strategic Stability Effects

Figure 1.5. Direct Strategic Stability Risks from AWS




Increased autonomy in weapon systems could also directly affect strategic stability and increase the danger of unintended escalation. Stockholm International Peace Institute (SIPRI) researchers have outlined various ways that AI and autonomous weapon systems may undermine strategic stability in their report *Artificial Intelligence, Strategic Stability, and Nuclear Risk*.[54] This includes:

- The possibility of autonomous remote sensing and distributed searching "undermining deterrence at sea;"[55]
- AI-enabled "left of launch" (i.e. in the systems that operate prior to a nuclear launch), and "autonomous cyber weapons;"[56]

[54] Vincent Boulanin, "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI,* 158, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf
[55] Ibid., 105. Note, however, that other scholars disagree, and find the possibility of "transparent oceans" unlikely except at certain geographic chokepoints, given the expense of UUVs. See Horowitz et al., "A Stable Nuclear Future?" 28.
[56] Ibid., 107.

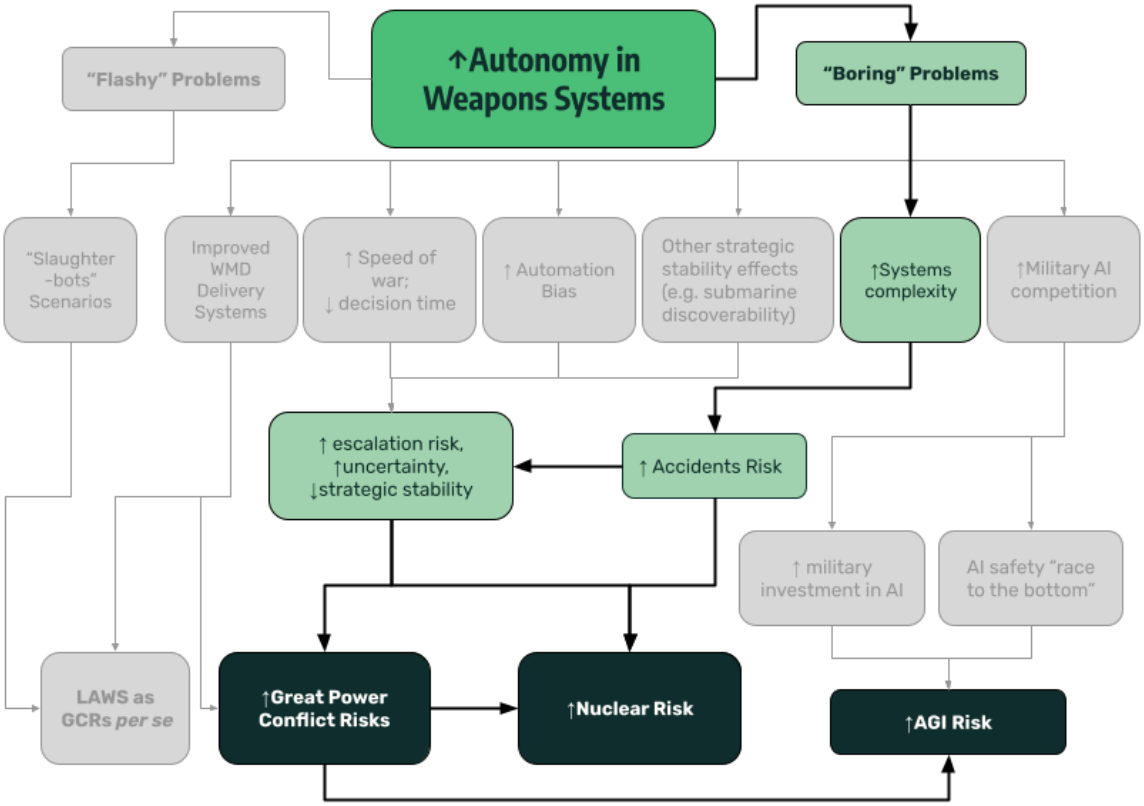
- Destabilizing AI-enabled/autonomous missile defense;[57]
- AI-enabled advances triggering sudden changes in states' nuclear policies and doctrines;[58]
- Automate parts of nuclear launch policy.[59]

While some of these scenarios may be more likely than others, several were discussed as possibilities by experts at SIPRI workshops, and all could be highly destabilizing.

[57] Ibid., 108.
[58] Ibid., 111.
[59] Ibid., 112.

## Pathway VI: Systems Complexity

### Figure 1.6. Systems Complexity and Accidents




Sixth, the increasing complexity of autonomous weapon systems may lead to an increasing risk of accidents. As Horowitz has explained, "normal accidents theory" of system failure "suggests that as system complexity increases, the risk of accidents increases as well and that some level of accidents are inevitable in complex systems," and that the risk of accidents in autonomous weapon systems may therefore be higher.[60] Proponents of normal accidents theory point to nuclear power plant and space mission accidents as historical evidence for the theory.[61] Sociological theories about normal accidents and black swans are not necessary, however, to understand how the complexity and "black box" character of some modern AI systems can lead to unexpected accidents. The complexity of machine learning algorithms has recently created unexpected behavior, especially when two autonomous systems interact. One example commonly cited in the literature on AWS

[60] Michael C. Horowitz, "When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability," *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 764–88, https://doi.org/10.1080/01402390.2019.1621174. See also Stephanie Carvin, "Normal Autonomous Accidents: What Happens When Killer Robots Fail?" (March 1, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3161446.
[61] See Charles Perrow, *Normal Accidents: Living with High Risk Technologies*.



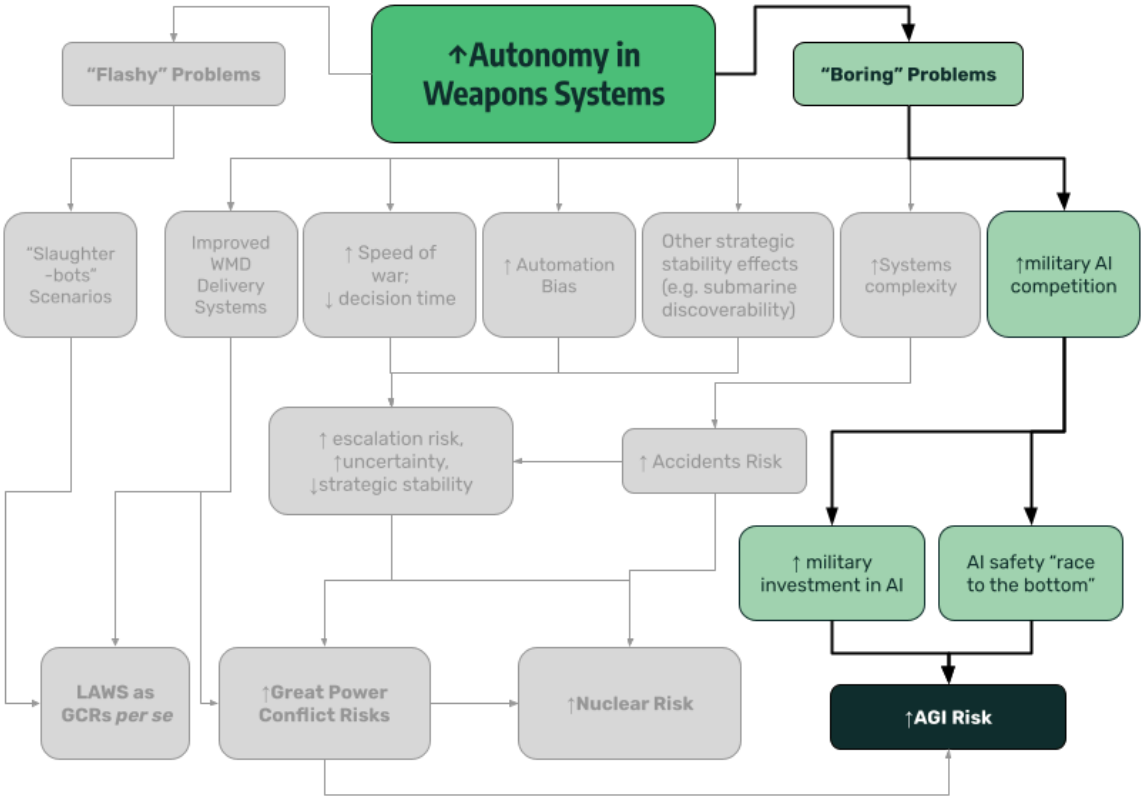
includes the case where price-setting algorithms on Amazon interacted to drive the price of an old biology textbook to nearly $24 Million.[62]

Such algorithmic mayhem and unexpected escalation is humorous with biology textbooks, but would be disastrous with autonomous systems in war that engage in similar rapid escalation in a wartime scenario. For example, one drone swarm might suddenly misinterpret an adversary swarm's micro-movements as "possibly hostile" in an unexpected way, assume a slightly more defensive formation, which the second swarm might then interpret as "possibly hostile," assume a more hostile position in turn, only confirming the first swarm's interpretation as "hostile," and so on, spiraling in this misinterpretation-action-misinterpretation pattern towards escalation. A similar dynamic can occur on other levels of conflict, including states' strategic decisions aided by AI-enabled decision-support systems.[63] Such unintended escalation could in turn lead to a shootdown that humans would then interpret as aggression and cause for further escalation. During the Cold War, for example, the shootdown of Korean Air Lines flight 007, which had strayed from its original flight path and was allegedly misidentified as a spy plane, caused one of the most tense moments between the two nuclear-armed superpowers.[64]

[62] Altmann and Sauer, "Autonomous Weapon Systems and Strategic Stability," 129.
[63] Thanks to Matt Lerner for pointing this out.
[64] For a discussion of KAL 007, see David Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*, 80-84.

# Pathway VII: Military AI Competition

## Figure 1.7. AWS and AI Competition




A final pathway towards existential risk runs through the potential for military AI competition. Several scholars have argued that the difficulty of verifying whether a system is "truly" autonomous, may make "AI arms races" more likely.[65] Many object to the term "AI arms race" (in part because AI is a general purpose technology, not a weapon) and this report will instead refer to "military AI competition."[66] In the words of AI scholar and regulator Missy Cummings, a "fake-it-till-you-make-it" attitude is prevalent in both commercial AI and the militarization of AI.[67] States have incentives — such as wanting to seem technologically advanced to adversaries and domestic audiences alike — to "fake it," or pretend a system has higher degrees of autonomy than it really does. Because AI is software-, not hardware-based, however, "it is not obvious in any given AI demonstration whether the results are real, or whether teams of humans are actually driving the success in a 'Wizard of Oz'

[65] Altmann and Sauer, "Autonomous Weapon Systems and Strategic Stability," 118.
[66] Paul Scharre, "Debunking the AI Arms Race Theory," *Texas National Security Review*, June 28, 2021, https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/.
[67] M. L. Cummings, "The AI that Wasn't There: Global Order and the (Mis)Perception of Powerful AI," *Texas National Security Review*, June 2020, https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/.

fashion."[68] This knowledge about effects extends to the operator of new software-based weapons as well, as demonstrated by Stuxnet and its code proliferation.[69] Thus, a state may interpret an adversary's "faked" autonomous system as evidence of real superiority, and as a result decide to invest more heavily in military AI itself. In this way, one state may trigger arms-race-like dynamics, *without possessing any meaningful AWS capabilities itself.* The threat of such competition thus becomes divorced from actual technological developments; whether the threat is real or perceived, destabilizing competition may ensue. Examples from the Cold War again confirm this dynamic. When President Reagan announced the pursuit of a non-existent capability of missile defense through the Strategic Defense Initiative (SDI or "Star Wars"), he appeared to hope that it would make nuclear weapons "obsolete," but instead confused Soviet planners and led to tension and near-arms races.[70]

Although the U.S. has publicly said that it will avoid delegating decision-making to machines, defense officials have also been clear that adversary behavior may force it to reassess this decision. For example, former Deputy Secretary of Defense Robert Work has hinted at such reassessment when he said that "We might be going up against a competitor who is more willing to delegate authority to machines than we are and, as that competition unfolds, we'll have to make decisions on how we can best compete."[71]

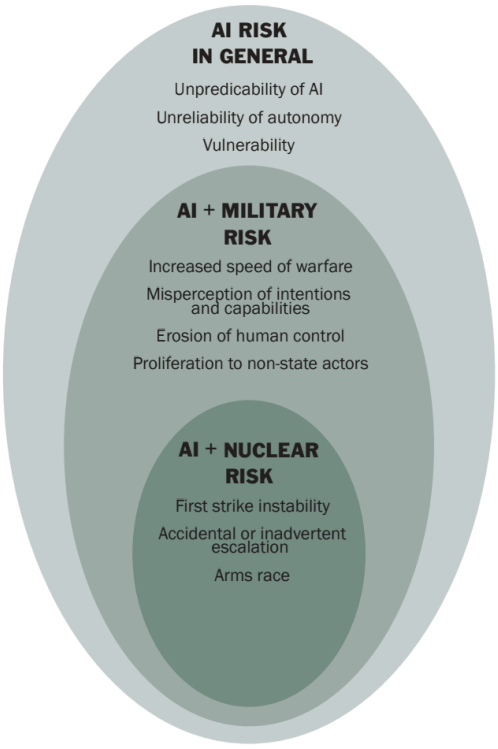
There are two ways that AWS-related military competition might trigger AI-related existential risk. First, it could accelerate the timelines of AGI development — if the U.S. and China both decided to pour many billions of dollars into the development of AI-enabled military systems, we might expect progress in AI to quicken, for better or worse. Second, such a dynamic might take the form of a "race to the bottom," where perception of competition with an adversary leads states to neglect AWS and AI safety.[72] Militaries have a history of disregarding accident risk for quick deployment. When the U.S. Department of Defense continued deployment of the V-22 Osprey despite its high rate of crashes and accidents during testing, an official explained, "Meeting a funding deadline was more important than making sure we'd done all the testing we could."[73] If competitive dynamics, combined with high accident tolerance, lead to (1) a higher likelihood of AGI development in the coming century and (2) a lower focus on safety, then the risks from bad AGI (see Founders Pledge's *Safeguarding the Future* report) might increase.

[68] Ibid.
[69] Thanks to Shaan Shaikh for pointing out this analogy.
[70] For a full discussion of SDI and the USSR's responses, see Hoffman, *The Dead Hand*.
[71] Robert O. Work, quoted in Altmann and Sauer, "Autonomous Weapon Systems and Strategic Stability," 124.
[72] Michael C. Horowitz, Lauren Kahn, and Christian Ruhl, "Introduction: Artificial Intelligence and International Security," *Texas National Security Review*, June 2020, https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/.
[73] "Saving the Pentagon's Killer Chopper-Plane," WIRED. Cited in Scharre, "Debunking the AI Arms Race Theory," 126.

# Unpacking the Risk

The previous sections showed that there are a variety of pathways through which increasing autonomy in weapon systems can increase global catastrophic risks. We can also understand this risk through a schematic developed by the Stockholm International Peace Research Institute (SIPRI), as reproduced in figure 2.

Figure 2: The Autonomous Risk Onion




*Source:* SIPRI, *AI, Strategic Stability, and Nuclear Risk,* 124.

Thus, we turn to the idea of "nuclear close calls" to conceptualize the possible risk of autonomous weapon systems.

## Nuclear Close Calls

How bad are these risks, and how can we better understand them? Each pathway outlined above poses its own risks, but one way to think about some rough quantification of the risk is how many nuclear "close calls" might have escalated to nuclear use in a AWS world vs. a non-AWS world. The Future of Life Institute has a timeline of nuclear "close calls". We can think of these as occurring "just below" the threshold of nuclear use, as indicated by the peak labeled "close call" in the yellow

line below — we have been very lucky. In a world with widespread use of autonomous weapons and the risks that they bring, we can imagine these close calls being "pushed" above the threshold of use.

In a world with the AWS risk factors identified above, how many close calls that we know about might have been "pushed over the edge" to nuclear use? From the Future of Life Institute list, there are 5 plausible candidates that involve one or more of the risk factors identified above:

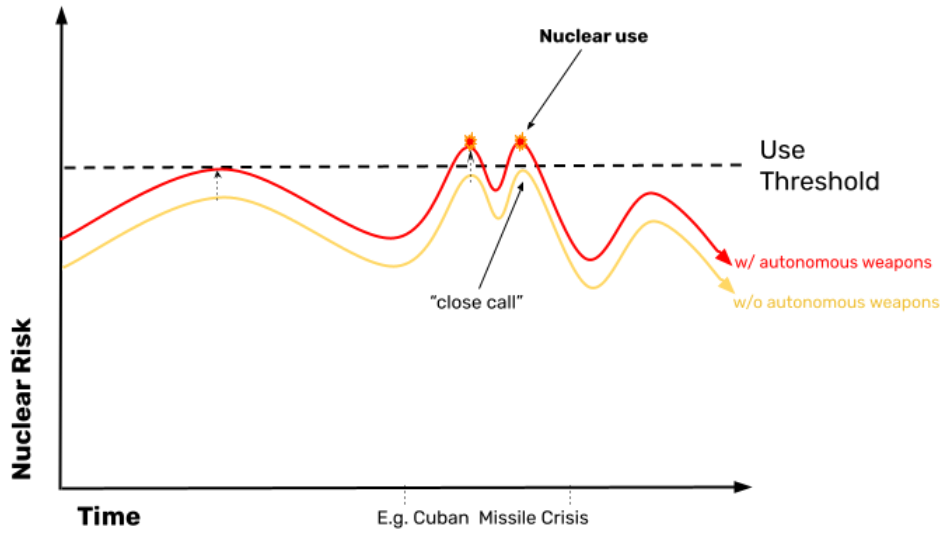
- **October 27, 1962**
  - Cuba shot down a U2 spy plane, an action U.S. leaders had previously agreed would automatically warrant launching an attack.[74] U.S. leaders changed their mind. Had more systems been automated, AWS might have automatically retaliated (or there may have been less time for human judgment), leading to escalation.
- **October 28, 1962**
  - Operators misidentified a satellite as incoming missiles over Georgia. "Fortunately, the reaction was ... slow enough that they were able to confirm no attack had actually occurred."[75] Increased speed of war in a AWS world might have reduced this time, and could thereby have led to nuclear use.
- **October 24, 1973**
  - A false alarm during DEFCON 3 led crews in Michigan to ready their bombers. "Pilots and crew all ran out to their B-52 bombers, ready to take off, when the duty officer realized it was a false alarm, and called them all back, before any further damage was done."[76] With autonomous bombers, and in a "hyperwar" world, there may have been much less time for recall.
- **November 9, 1979**
  - "Computers at NORAD headquarters indicated a large-scale Soviet attack on the United States" because a technician had accidentally inserted a training tape with a simulated attack.[77] Officials may not have had six minutes to wait for corroborating evidence in a "hyperwar" world where war happens much more quickly.
- **September 26, 1983**
  - "Stanislav Petrov, the Soviet officer on duty, had only minutes to decide whether or not the satellite data were a false alarm."[78] Increased speed of war and automation bias in a world with widely-integrated AI-enabled systems may have brought the alarm to Soviet leadership, increasing the risks of a nuclear strike (see sections III and IV, above)

[74] Future of Life Institute, "Accidental Nuclear War: A Timeline of Close Calls," Future of Life, https://futureoflife.org/background/nuclear-close-calls-a-timeline/.
[75] Ibid.
[76] Ibid.
[77] Ibid.
[78] Ibid.

Even if only 50% of these close calls were "pushed over the edge," this would represent a **2.5-fold increase in nuclear use** between the non-AWS world we knew (where only Hiroshima/Nagasaki happened) and the hypothetical AWS world. This is an explicitly speculative approach, and we emphasize that this explanation is intended not to actually quantify risk, but to illustrate it and to formalize our theory of change. History shows that nuclear close calls are frequent and have come frighteningly close to nuclear use. This exercise is intended to illustrate that tipping the balance of risk by altering the character of war through AI-enabled warfare, therefore, could have dramatic — and possibly catastrophic — consequences.

We believe that being explicit in our model of AWS risk can help to lay bare any flaws and improve the quality of our assessments. It is important to note, moreover, that the preceding section illustrates the increased risk of **just one of many pathways**, some of which might be relatively independent (e.g. short-term nuclear war risk and long-term AGI risk).

Figure 3. Autonomy and Nuclear "Close Calls"




## Setting Precedents for AI Regulation

In addition to the risks discussed above, we believe that decisions related to the development and regulation of AI-enabled weapons systems may affect the future of powerful artificial intelligence, including general intelligence.[79] As we discuss in the Executive Summary of our Safeguarding the Future report, "Managing the transition to AI systems that surpass humans at all tasks is likely to be

[79] Thanks to external reviewers Anthony Aguirre, Stephen Clare, and Emilia Javorsky for pointing out the need to discuss general AI risks.

one of humanity's most important challenges this century, because the outcome could be extremely good or extremely bad for our species." The risks discussed in this report arise from existing narrow AI, but we believe it is plausible that progress on regulating these risks may help lay the groundwork for more effective management of long-term AI risks.

## The Funding Landscape: Neglectedness and Diversification

So far, we have argued that autonomous weapon systems are an important cause area, one that has the potential to cause great harm to humanity, and possibly lead to human extinction. An important problem for humanity, however, is not *necessarily* a good problem area for philanthropists, for example if the problem is not neglected and potential interventions are well-funded. In this case, we believe that the risks from autonomous weapon systems are highly neglected by philanthropists, and outline the funding landscape in this section.

First, autonomous weapons risks are neglected relative to government spending on research and development. That is,  Compare, for example, autonomous weapons with another militarized technology: cyber capabilities. From interviews with experts, we believe that the total philanthropic funding related to autonomous weapons is around $2.5 million a year, though we are uncertain about this number.[80] As discussed above, according to the Center for Security and Emerging Technology, the FY 2021 U.S. Defense Budget requested $1.7 *Billion* of investments for "autonomy to enhance 'speed of maneuver and lethality in contested environments' and the development of 'human/machine teaming,'" in addition to other military applications of AI.[81] Thus, the ratio of *global* philanthropic funding to military investments in the U.S. alone is 1:680. In cybersecurity, by contrast, philanthropists (such as the Ford Foundation, the Hewlett Foundation, and Cisco's corporate giving program) spent $159 Million on cybersecurity in 2021, compared to the DoD's $9.8 Billion cyber spending, a less extreme ratio of 1:62.[82] This does not even include growing spending by private

[80] Several organizations are working on autonomous weapons risks, including Stop Killer Robots (a coalition of NGOs), the Convention on Certain Conventional Weapons Group of Governmental Experts (a UN working group), the International Committee of the Red Cross, the Future of Life Institute, the Center for a New American Security, and others. The majority of this funding is through the NGO coalition Stop Killer Robots. The rest of the funding is smaller grants, as outlined in Candid's Peace and Security Funding index. To illustrate these smaller grants: the Smith Richardson Foundation gave CNAS $125,000 in 2017 for related work; the Stanley Center for Peace and Security gave $102,825 to an anonymous recipient in 2017, and several smaller grants that appear related, like a $34,750 grant in 2017 for "Commissioned policy analysis and policy salon-style dinners focused on the intersection of emerging technology and nuclear weapons policy"; the Joseph Rowntree Charitable Trust made a 2017 grant of $26,214 and a 2015 grant of $17,188; the Samuel Rubin Foundation made two grants in 2015 and 2016, each for $10,000 to Human Rights Watch, to support the Campaign to Stop Killer Robots.

[81] "U.S. Military Investments in Autonomy and AI: A Strategic Assessment," *Center for Security and Emerging Technology*, https://cset.georgetown.edu/publication/u-s-military-investments-in-autonomy-and-ai-a-strategic-assessment/.

[82] Search on Candid Foundation Maps; https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2021/fy2021_Budget_Request_Overview_Book.pdf; 1-5

cybersecurity firms, which would likely make the difference look even larger.[83] In other words, when comparing philanthropic spending on autonomy in weapon systems to cyber capabilities, autonomous weapons are under-funded by more than a factor of 10. This is especially concerning given the potential risks outlined above.

Moreover, regardless of whether autonomous weapon systems are philanthropically neglected in general, the *strategic* risks of autonomous weapons and militarized AI are especially neglected. From our research, we have found that most civil society organizations focus on humanitarian and ethical issues. In the words of one scholar, "The fact that the issue's framing has been dominated by NGOs campaigning to ban 'killer robots' affects the debate. Potential harm to civilians has been front and center in the discussion. Strategic issues … have taken a back seat."[84] We would note that potential harm to civilians should remain a central issue, but that strategic issues like nuclear stability seem especially likely to create such harm on a massive scale. Thus, philanthropists' focus should shift towards strategic risks.

This matters because the pathways to global catastrophic risks identified above are not humanitarian, but strategic, focusing on great power war, unintended escalation, and nuclear stability. We therefore conclude — although the work on humanitarian issues is important and praiseworthy — that the allocation of philanthropic funds is underweighted on the strategic risks of autonomous weapons. The risks of this report are therefore doubly neglected; once because the issue area as a whole is neglected, and once because the most consequential risks are neglected within it. Effective philanthropists could help to balance this misallocation, and we can therefore deduce the first guiding principle for mitigating risks from autonomy in weapon systems: **focus on strategic risks.**

[83] Thanks to Shaan Shaikh for pointing this out.
[84] Scharre, *Army of None,* 348.

# Evaluating Interventions

To summarize the preceding pages: increasing autonomy and applications of AI in the military realm is an important problem that could exacerbate existential and global catastrophic threats through several pathways, such as speed and automation bias. Moreover, this problem is neglected by philanthropists, and the little funding that does flow is focused on humanitarian, not strategic, risks. Knowing that the problem is both important and neglected, this section asks: What can we do? Are there tractable interventions with evidence for success?

This section is divided into three sub-sections. First, we discuss the problems with formal treaty-based arms control, often framed as a "killer robots ban." Second, we explain why, when evaluating interventions, we should prioritize those that minimize the need for international coordination.

## The Problems with a "Killer Robot Ban"

Intuitively, there is a simple solution to the problem of autonomous weapons: ban them. In this section, we outline many of the problems with a full ban, and conclude that effective philanthropists should instead focus on other interventions.

Many of the potential problems with a formal legal treaty or "ban" on killer robots have been discussed in depth elsewhere, especially in Paul Scharre's *Army of None*. Here, we briefly list these problems with footnotes to allow the reader to find more information:

1. **Bans have a high failure rate**: Humans have tried to "ban" many instruments of war — including crossbows, airplanes, and submarines — with very limited success. Even when bans are passed (often when major militaries see little utility in a class of weapons), they are often violated, as in the case of chemical weapons.[85]
2. **A ban on software is difficult to verify**: Fundamentally, what distinguishes an autonomous weapon from a remotely-piloted drone is software; the two systems "look" the same, and may behave the same, posing great difficulties to treaty-based verification.[86]
3. **The line between autonomous and automated systems is blurry**: The UN body tasked with discussing autonomous weapon systems has spent the last eight years mostly on a definitional debate, suggesting that the very blurriness of the problem might preclude progress on a ban.[87]

[85] See chapter 20, "The Pope and the Crossbow," in Scharre, *Army of None,* 331-345.
[86] "The difference between a remotely piloted system and an autonomous system is software, not hardware, meaning verification that a given country is operating an autonomous system at all would be difficult." Horowitz, "When Speed Kills," 780; see also Scharre, *Army of None,* 344.
[87] Scharre, *Army of None,* 346.

4. **AI is everywhere:** Unlike some banned military technologies, like landmines, AI is a "general purpose technology" being integrated by commercial and other private actors throughout modern life. This makes controlling its use more challenging.[88]
5. **Autonomous systems provide military advantages:** Powerful militaries view autonomy and the applications of AI to the battlefield as militarily advantageous; historically, bans have worked best on weapons that are not seen as decisive.[89] Therefore…
6. **The most powerful states oppose a ban**: The world's most powerful militaries — and those states with the most institutional power in the UN — oppose the idea of a full ban on autonomous weapon systems.[90]

Related to the final point, this opposition to a ban is not limited to the United States. Fu Ying, China's former Vice Minister for Foreign Affairs, has written, "Some experts and scholars from around the world simply advocate for a blanket ban on developing intelligent weapons that can autonomously identify and kill human targets, arguing that such weapons should not be allowed to hold power over human life…. As things stand, however, a global consensus on a total ban on AI weapons will be hard to reach, and discussions and negotiations, if any, will be protracted."[91]

This does not necessarily mean that a ban would be harmful or completely useless. We admire some of the civil society actors advocating for a ban, and appreciate arguments around normative change and creating stigma.[92] For the reasons listed above and the funding imbalance discussed in the previous section, however, we do not believe that support for a ban is the most effective use of marginal philanthropic funds.

## Focus on Key Stakeholders, Not Multilateralism

In its simplest form, the question of whether to pursue a legally binding treaty on autonomous weapon systems is a coordination problem, which becomes more challenging as more countries join. As Scharre puts the problem:

> "The number of countries that need to participate for a ban to succeed is also a critical factor. Arms control was easier during the Cold War when there were only two great powers. It was far more difficult in the early twentieth century, when all powers needed to agree. A

[88] Horowitz, "When Speed Kills," 777.
[89] Scharre, *Army of None,* 332, 340.
[90] See, e.g. the statements made by a U.S. official at the latest CCW meeting, which also point towards Great Power appetite for confidence-building measures rather than a formal ban, as explained below. "U.S. Rejects Calls for Regulating or Banning Killer Robots," *The Guardian,* https://www.theguardian.com/us-news/2021/dec/02/us-rejects-calls-regulating-banning-killer-robots. (Technically, the headline is inaccurate, as the U.S. has called for a "code of conduct," as acknowledged in the article.)
[91] Fu Ying, quoted in "The U.S. and China Can Stop AI Warfare Before It Starts," *Noema*, https://www.noemamag.com/the-u-s-and-china-can-stop-ai-warfare-before-it-starts/.
[92] Thanks to Anthony Aguirre and Emilia Javorsky for pointing to this argument.



> single defector could cause an arms control agreement to unravel. Since the end of the Cold War, this dynamic has begun to reemerge."[93]

The discussions on autonomous weapons take place within the Convention on Certain Conventional Weapons (CCW), which operates by consensus, another block to coordination. In the words of one researcher at Human Rights Watch, "The CCW, unlike some disarmament forums, operates by consensus, which allows a single country to block a proposal."[94]

We have used similar reasoning before at Founders Pledge when evaluating climate charities, where we prioritized solutions with a "limited need for cooperation."[95] When comparing different interventions for autonomous weapons risk, too, we can place greater weight on interventions that require fewer actors to coordinate. Thus we can state the second guiding principle: **minimize the need for multilateral coordination**, and prioritize those actors that matter most to the most pressing risks. We believe these actors include the great powers, especially the United States and China, as well as potentially some other states involved in the proliferation of relevant hardware.[96]

In short, the more actors need to coordinate for a solution, the more challenging coordination becomes. This is especially relevant, as we have seen above that much of the risk landscape flows only through the great powers and other nuclear states; their coordination matters most to mitigating catastrophic risks.

What *can* philanthropists do? After reviewing the evidence on this cause area, we suggest two broad classes of interventions: research on strategic risks and work on confidence-building measures.

## Funding Research on Strategic Risks of AWS

As noted above, strategic issues remain neglected in the discussions of AI-enabled military systems.[97] Understandably, immediate issues about the ethics, responsibility, the laws of war, human dignity, and the inhumanity of new killing systems have been the focus of much scholarship, NGO work, and discussions at the CCW. We believe, however, that the *scale* of suffering wrought by potential strategic risks — such as large-scale nuclear war — would be far greater. As discussed above, however, a comparatively small amount of research has focused on this issue, which is therefore especially neglected relative to its importance. Therefore, our first candidate intervention

[93] Scharre, *Army of None,* 340.
[94] "Countering Consensus through Humanitarian Disarmament: Incendiary Weapons and Killer Robots," *Human Rights Watch* (blog), December 21, 2021, https://www.hrw.org/news/2021/12/21/countering-consensus-through-humanitarian-disarmament-incendiary-weapons-and-killer.
[95] "Climate Change Executive Summary," Founders Pledge, https://founderspledge.com/stories/climate-change-executive-summary.
[96] Thanks to Emilia Javorsky for pointing to the importance of proliferator-states.
[97] Scharre, *Army of None,* 348.

is funding further research on the strategic risks of autonomous weapon systems (and of near-term military applications of AI in general).

We are especially interested in interventions that increase the *quality of evidence* surrounding arguments for and against AWS risks. Much of the evidence in the studies cited above relies on expert opinion, reasoning by analogy, and historical comparisons. Of course, randomized controlled trials are not an option with such dangerous technology. Nonetheless, quasi-experimental methods have recently emerged in the field of International Relations (IR). IR and foreign policy analysis has relied on "wargaming" and scenario-planning techniques since the early Cold War.[98] Recent work on "experimental wargames" — randomly assigning participants to groups and varying the wargame conditions to assess changes in decision-making and reasoning — has advanced this methodology. Combined with advances in computer simulation, these techniques have led to the adoption of more rigorous methods that can help create "human-derived, large-*n* datasets for replicable, quantitative analysis" — in other words, higher-quality evidence.[99]

Philanthropists can help to fund such experimental wargames surrounding major open questions on AWS, including but not limited to: How bad would automation bias be in battlefield situations with autonomous weapons? How does the speed of (simulated) war affect the quality of decisions when AWS are involved? Is escalation more or less likely when the systems involved are autonomous systems? How much credence to participants place in unsubstantiated claims that an adversary deployed lethal autonomous systems?

Specifically, philanthropic funding could pay for researcher time, participant time and expenses (especially if participants are high-level decision-makers), designing software for crisis simulation, paying wargame designers or consultants, and more.

Additional research could focus specifically on how AI-enabled military systems affect nuclear stability. Despite some promising early research on the impact of autonomous systems on nuclear stability, we believe more work would help to stress-test these ideas, uncover new pathways for risk, and create a clearer understanding of the tractability of this problem. We can conceptualize the cost-effectiveness of such research as the decision-making value of information. This value of information is especially clear when understood through the lens of funding decisions. Future funding decisions about AWS — by philanthropic organizations and by governments — could be on the order of tens or hundreds of millions of dollars. If new research were to show that AWS are highly unlikely to be a large risk, for example, then this philanthropic and government money could be diverted from AWS work and spent on other more proven effective interventions. We can therefore consider the cost-effectiveness of such research such that the counterfactual philanthropic allocation is equal to the change in allocation between a world with this research and a world

[98] Sharom Ghamari-Tabrizi, "Simulating the Unthinkable: Gaming Future War in the 1950s and 1960s," *Social Studies of Science 30,* no. 2, (2000), https://www.jstor.org/stable/285834; for a recent overview of experimental wargames as a research tool, see Erik Lin-Greenberg, Reid B.C. Pauly, and Jacquelyn G. Schneider, "Wargaming for International Relations Research," *European Journal for International Relations 28,* no 1, (2021), *https://journals.sagepub.com/doi/abs/10.1177/1354066121106409O?journalCode=ejta*.
[99] Andrew Reddie et al., "Next Generation Wargames," *Science,* 362, no. 6421 (2018), https://www.science.org/doi/10.1126/science.aav2135.



without it. The cost-effectiveness is that difference divided by the money spent on the research. As a rough illustration, if 100K spent on AWS research were to precipitate a $10 Million allocation change away from totally-ineffective AWS-mitigation efforts and towards proven Malaria-prevention charities, and the cost of averting a death through such charities is around $5,000, then the counterfactual value of this research would be: $\frac{\$10{,}000{,}000 \times (\frac{1\ Life}{\$5{,}000})}{\$100{,}000}$, or ~$50 per life saved — a highly effective intervention. More realistically, rational philanthropists will not make such a large update to their beliefs about AWS-mitigation effectiveness, but will reason probabilistically: if a high-quality study leads to a 10% belief update, and therefore a 10% allocation change, then the cost-effectiveness in the hypothetical world be $500 per life saved — still highly effective. Even just a 1% belief update would still put the effectiveness of this charity on the same order of magnitude on cost-effectiveness as highly-effective Malaria-prevention charities themselves. Additionally, the counterfactual giving would likely be far more complicated, likely involving funging of money not with current-generations global health interventions, but with other longtermist opportunities. Nonetheless, this reasoning would still apply.

Moreover funding research into mitigation, such as uncovering new confidence-building measures (CBMs), would also have a high expected value, as each new promising mitigation effort that is uncovered will raise the probability that mitigation overall will be successful. In addition to uncovering new CBMs, such research could also refine existing proposals and study the obstacles to their implementation.[100] Below, we list several distinct AWS/military AI CBMs that have been proposed thus far, in just a small handful of research papers. A concerted research effort could increase this number, thereby increasing the probability that any one of the proposed CBMs will be adopted and successful. Finally, we believe a promising avenue for research would explore what strategies and policies would increase or decrease the tractability of arms control of AI-enabled weapons systems in the long run, especially for systems that affect strategic stability. We believe that formal arms control (like a "Killer Robot Ban") remains intractable at the moment for the reasons above, but we are open to the possibility that well-targeted policies could shift the balance in favor of arms control in the long run.

## Field-Building

As a subset of "funding research on strategic risks," we view field-building as a potentially valuable intervention in this space. The field of researchers who study the issues outlined in this report is small. For example, we identified 53 think tanks and other organizations whose work is plausibly relevant to autonomous weapons research, but only 9 who work actively on strategic risks from autonomy. We believe the neglectedness of research on strategic risks could therefore be addressed especially well through field-building interventions that foster expertise and future researchers.

When possible, therefore we would recommend field-building interventions, like helping to start a new organization that focuses on emerging technology security risks, as a valuable type of intervention. However, such interventions can be challenging for individual donors, as we explained

[100] Thanks to Carl Robichaud for pointing this out during external review.

in our Great Power Conflict report: "field-building requires large amounts of capital and long-term commitments, which means it is likely not the optimal intervention for individual donors to consider in this space."[101]

## Confidence-Building Measures

Recent work on regulating autonomous weapon systems has identified one promising candidate that can fulfill these three guiding principles: confidence-building measures, or CBMs.[102] The UN Office of Disarmament Affairs defines CBMs as "planned procedures to prevent hostilities, to avert escalation, to reduce military tension, and to build mutual trust between countries."[103]

For example, after a lengthy treatment of the nuclear risks of autonomy and AI, researchers at SIPRI concluded, "CBMs provide a useful compromise [between unilateral measures and hard-to-negotiate international law] in this regard. They would not need to be legally binding [...], but they would be easier to reach agreement on."[104] The importance of confidence-building and transparency for autonomous weapons regulation is also discussed in a recent roadmapping exercise on autonomous weapons regulation.[105] Below, we explain in detail why we agree that CBMs are among the most promising interventions, especially when compared to current frameworks of "banning" autonomous weapons.

For example, the Helsinki Final Act's CBMs included measures like "voluntary reporting with at least a 21-day prior notification of military maneuvers that would exceed over 25,000 troops and that would occur within 250 kilometers from a state's border," designed to reduce uncertainty and the potential for unintended escalation.[106] Unlike formal arms control, which legally prohibits the development or use of certain kinds of weapons, confidence-building measures are agreements designed to reduce risk in a way that all parties consider to be in their interest, often through the tools of "communication, constraint, transparency, and verification."[107] The "nuclear hotline"

---

[101] Stephen Clare, "Great Power Conflict," *Founders Pledge,* 112, https://assets.ctfassets.net/x5sq5djrgbwu/1cT1AsYClomlUh0que0JgW/74ffb0b1a4803e1f9654e86e2c7f3af1/Great_Power_Conflict_report_-_Founders_Pledge.pdf.

[102] 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 (January 1, 2020), https://doi.org/10.1016/j.orbis.2020.08.003; and CNAS, "AI and International Stability: Risks and Confidence-Building Measures," https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[103] United Nations, "Military Confidence-Building – UNODA," https://www.un.org/disarmament/cbms/.

[104] Vincent Boulanin, "Artificial Intelligence, Strategic Stability and Nuclear Risk," *SIPRI,* 158, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf

[105] Ronald Arkin et al., "A Path Towards Reasonable Autonomous Weapons Regulation," IEEE Spectrum, https://spectrum.ieee.org/a-path-towards-reasonable-autonomous-weapons-regulation.

[106] Erica D. Borghard and Shawn W. Lonergan, "Confidence Building Measures for the Cyber Domain," *Strategic Studies Quarterly* 12, no. 3 (2018): 14.

[107] "Confidence-Building and Nuclear Risk-Reduction Measures in South Asia • Stimson Center," *Stimson Center* , June 14, 2012, https://www.stimson.org/2012/confidence-building-and-nuclear-risk-reduction-measures-south-asia/.

established between the U.S. and U.S.S.R. after the Cuban Missile crisis exemplifies one such CBM, designed to share information, communicate during crises, and avoid misunderstandings and unintended escalation.[108] We have previously recommended another well-known kind of confidence-building measure — track II dialogues — to address risks from great power conflict.

Figure 5. Comparison of Ban and CBMs

| | Formal Ban Treaty | Confidence-Building Measures |
|---|---|---|
| **Strategic Focus** | Diffuse – most of the conversation focuses on international humanitarian law and human rights. | Targeted – CBMs are specifically designed to increase strategic stability. |
| **Required Multilateral Coordination** | High — the CCW requires consensus, and other treaties also require many states parties to agree. | Flexible – CBMs can be just between two states, but can easily be expanded to more states. |
| **Tractability** | Low — no results after years of discussion; powerful states oppose this solution. | Medium — powerful states have expressed a preference for this solution, and broad multilateral coordination is not necessary; however, all international coordination is difficult. |

Confidence-building measures were prominent during the Cold War, when nuclear strategists similarly sought interventions to increase international stability, especially after the Cuban Missile Crisis. The focus was strategic risks of escalation, the measures were bilateral — between the superpowers — and they preserved the theorized benefits of nuclear deterrence. As Scharre and Horowitz write, "While both sides recognized that war might occur, they had a shared interest, due to the potentially world-ending consequences of a global nuclear war, in ensuring that any such outbreak would be due to a deliberate decision, rather than an accident or a misunderstanding."[109]

The United Nations Office of Disarmament Affairs has compiled a Repository of Military Confidence Building Measures, including broad measures such as:

- "Establish a direct communications system, or "hotline" between heads of state, ministers of defence, chiefs of military forces, and/or military commanders"

[108] 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 (January 1, 2020): 537, https://doi.org/10.1016/j.orbis.2020.08.003.
[109] Scharre and Horowitz, *AI and International Stability,* 11.



- "Exchange information on relevant policies, doctrines and tactics, defence policy papers and national legislations."
- "Hold regular meetings of military officials to exchange information, and discuss common operational issues and concerns"
- "Establish military research contacts and collaboration"
- "Agree to exchange invitations to observe demonstrations of new weapon systems"
- "Establish joint crisis-management / conflict prevention centres"
- and many more.

More specific confidence-building measures can be applied to military AI and autonomous weapons, as explained in the next section.

## Actionable CBM Options

The most comprehensive list of AI-CBMs applicable to autonomous weapon systems is outlined in a recent CNAS report on *AI and International Stability*:[110]

- Formulating and promoting "AI norms"
- Track II dialogues and military-military dialogues
- Creating an AI Code of Conduct
- Public signaling for the importance of rigorous testing and evaluation (T&E)
- Increased transparency about T&E
- International military AI T&E standards
- Shared AI safety civilian research
- "International Autonomous Incidents Agreement"
- Markers for autonomous systems
- Designating off-limits geographic areas

As well as nuclear-specific AI CBMs:

- Agreeing to "strict human control" for nuclear launch decisions[111] (N.B., verification concerns apply here as well)
- Prohibiting "uninhabited nuclear launch platforms"[112] (N.B. this is a ban-like intervention, and thus we believe some of the considerations discussed above apply)

Not all CBMs are created equal, however. Several of them share the twin problems of difficult verification and easy cheating. Consider, for example, the idea of a visual marker for autonomous systems. The idea is appealing for its simplicity and its familiarity, akin to the red cross, red crescent,

[110] The following list is from Scharre and Horowitz, *AI and International Stability,* 12-19.
[111] Ibid., 20.
[112] Ibid., 20-21.



or red crystal that is supposed to protect healthcare providers in war.[113] Medics, however, have little incentive for passing as soldiers and risk getting shot while providing care. There may, however, be incentives for marking an autonomous system as non-autonomous, such as surprising an adversary with swift and unexpected machine behavior. Verifying the difference, as explained above, is near-impossible because the hardware could look the same.

Below, we discuss two promising examples of confidence-building measures in greater detail: track II dialogues and an International Autonomous Incidents Agreement. Although we believe that these two issues are promising, other CBMs may turn out to be highly effective interventions as well.

## Track II Dialogues

At Founders Pledge, we have previously recommended the funding of track II dialogues between experts from the United States and China to mitigate the risks of great power conflict. As we noted in that report:

> "Track II diplomacy initiatives, which involve bringing together non-governmental representatives from two [or more] countries, like scientists or businesspeople, to share information and discuss mutual problems, appear highly neglected. Since there are theoretical and empirical reasons to believe that Track II diplomacy can play an important role in dispute resolution, especially when official diplomatic channels are strained or even fully cut off, we think this is a highly-promising intervention."

Further track II (and eventually track 1.5) dialogues on AI safety have been recommended by other organizations whose research we trust, including Georgetown University's Center for Security and Emerging Technology (CSET).[114] Dialogues specifically focused on autonomous weapon systems are a natural extension of this framework.

Another reason to fund track II dialogues to mitigate risks from autonomous weapons is that we believe these dialogues have a high leverage factor. In addition to their beneficial diplomatic effects, such dialogues can help to expand the search for promising new CBM candidates, and to evaluate their effectiveness more rigorously. A potential downside is that such dialogues may provide an opportunity for bad actors to spread misinformation (for example, a state ordering purportedly "independent" scientists to downplay the state's capabilities). Philanthropists concerned about this possibility may need to ask organizations about the invitation-vetting process. Philanthropists can fund the expenses of track II dialogues — inviting researchers and speakers, the costs of conferences, the cost of associated publications, related research, staff time, and more.

---

[113] Christian Ruhl, "Note to Nations: Stop Hacking Hospitals," *Foreign Policy* (https://foreignpolicy.com/2020/04/06/coronavirus-cyberattack-stop-hacking-hospitals-cyber-norms/).
[114] "AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement," *Center for Security and Emerging Technology*, https://cset.georgetown.edu/publication/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/.

## International Autonomous Incidents Agreement

An International Autonomous Incidents Agreement (IAIA) has historical precedent, potential great-power support, and can become a focal point for greater cooperation. Scharre and Horowitz base the idea of an IAIA on the 1972 Incidents at Sea Agreement (formally the "Agreement on the Prevention of Incidents on and over the High Seas"):

> "The Incidents at Sea Agreement, not initially considered a prominent part of the 1972 SALT I accord, created a mechanism for communication and information surrounding the movement of U.S. and Soviet naval vessels. The agreement regulated dangerous maneuvers and harassment of vessels, established means for communicating the presence of submarines and surface naval movements, and generated a mechanism for regular consultation. These successes helped lead to the formalization of the CBM concept in 1975 in Helsinki at the Conference on Security and Cooperation in Europe."[115]

Similarly, an IAIA could create a common understanding of "rules of the road" for autonomous weapons and communication channels and norms regarding the deployment of autonomous systems. The terms of such an agreement, Horowitz and Scharre explain, could be self-enforcing, "such that it is against one's own interests to violate them."[116]

The original Incidents at Sea Agreement was considered a success. As early as 1984, the U.S. Secretary of the Navy credited the Incidents at Sea Agreement with a decrease in collisions and a "businesslike" resolution of situations that previously would have become diplomatic crises — and that this stability continued even during the Soviet invasion of Afghanistan, an otherwise tense situation between the superpowers.[117] According to retired Rear Admiral Eric McVadon, the average number of incidents dropped after the agreement, from over 100 throughout the 1960s to around 40 in 1974, with *zero* incidents during the tense standoff of the 1973 Arab-Israeli War.[118] There is of course an endogeneity problem with such data; if geopolitical tensions were low enough to negotiate an agreement (1972, after all, was in a period of U.S.-Soviet *détente*), then the low geopolitical tensions may have made serious incidents at sea less likely, with or without an agreement. Nonetheless, later statements by officials of the Reagan Administration — during a time of heightened tension, after the Soviet invasion of Afghanistan, the 1980 Olympic boycott, the Strategic Defense Initiative, and the breakdown of initial START and INF negotiations — suggest that the decrease in incidents continued, and officials viewed the agreement as an important factor. In 1983, Secretary of the Navy John Lehman credited the agreement with improving Navy-to-Navy

[115] Scharre and Horowitz, *AI and International Stability,* 11.
[116] Ibid., 17.
[117] Sean M. Lynn-Jones, "A Quiet Success for Arms Control: Preventing Incidents at Sea," *International Security* 9, no. 4 (1985): 154–84, https://doi.org/10.2307/2538545.
[118] Eric A. McVadon, "The Reckless and the Resolute: Confrontation in the South China Sea," *China Security* 5, no. 2 (Spring 2009): 10.

communications and in 1985 noted that incidents were “way down” in the 1980s from the 1960s and early 1970s.[119]

The Agreement, moreover, became a model for other types of cooperation, including similar agreements between the U.S.S.R. and Japan and other U.S. allies, as well as military accords between Greece and Turkey, between Indonesia and Malaysia, and between the United States and China.[120] Scharre and Horowitz point to the historical legacy of this agreement as a potential factor for tractability: “Given the perceived success of the Incidents at Sea Agreement in decreasing the risk of accidental and inadvertent escalation between the United States and the Soviet Union, an equivalent agreement in the AI space might have potential to do the same for a new generation.”[121]

There is additional evidence to support the notion of tractability in this space, because senior military leaders from China have expressed support for such measures: “in a February 2020 article, Senior Colonel Zhou Bo in China’s People’s Liberation Army (PLA) advocated for CBMs between the United States and China, including on military AI, drawing on the example of the 1972 Incidents at Sea Agreement.”[122] In the article, Colonel Zhou points to the Cold War CBMs as a model for stability: “outright conflict was averted, thanks to a few modest agreements and well-established hotlines for emergency communication. Even bitter enemies can build trust, and with imperfect tools, when they measure the stakes of a full-on clash.”[123] Zhou also praised the success of the Incidents at Sea Agreement (“the agreement does seem to have drastically reduced the overall risk of dangerous encounters”) and explicitly called for stronger CBMs in new domains, including “space exploration, cyberspace and artificial intelligence.”[124]

What, specifically, could a philanthropist do to facilitate an International Autonomous Incidents Agreement?  Philanthropists could fund groups to work on the drafting of the details of the International Autonomous Incidents Agreement. Funding researchers, research assistants, publication expenses, and roundtables to discuss the contents of such an agreement are all within the purview of private philanthropists. Track II dialogues, moreover, could mature into official discussions that lead to such an agreement. Moreover, philanthropists could help disseminate the findings of this kind of work, funding public-relations campaigns, op-eds, and activist groups to bring ideas for an IAIA to governments and international organizations.

---

[119] U.S. Department of State (Archived Content), *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*: https://2009-2017.state.gov/t/isn/4791.htm.
[120] David F Winkler 1, “The Evolution and Significance of the 1972 Incidents at Sea Agreement,” *Journal of Strategic Studies* 28, no. 2 (April 1, 2005): 372, https://doi.org/10.1080/01402390500088395.
[121] Scharre and Horowitz, *AI and International Stability,* 17.
[122] Ibid.
[123] Zhou Bo, “Opinion: China and America Can Compete and Coexist,” *The New York Times*, https://www.nytimes.com/2020/02/03/opinion/pla-us-china-cold-war-military-sea.html.
[124] Ibid.

# Conclusion

Autonomous weapon systems and related military AI applications pose risks to the future of humanity and current generations through a variety of pathways, including as risk factors for great power conflict, nuclear exchange, and risks from misaligned general artificial intelligence. We believe that the current philanthropic landscape is lopsided in focusing excessively on regulation via a formal "ban" — an intervention with low tractability. Moreover, much current funding focuses on issues of ethics and international humanitarian concern, whereas some of the strategic risks outlined above have been neglected except by a few scholars and research groups. We suggest that effective philanthropists should focus on funding research on strategic risks, as well as on non-ban governance mechanisms, like confidence building measures. The risks from AI-enabled warfare are more insidious than "killer robots," and run through decision-support and early-warning systems, as well as other "boring" risks. Identifying and addressing these risks may be crucial to the future survival and flourishing of humanity.

Funding opportunity recommendations will be outlined in separate Founders Pledge publications.

# Disclosures

The author has previously worked with Professor Michael Horowitz, one of the experts whose work is cited frequently in this report. The author has also organized academic workshops on AI that have included some of the experts cited above. The author has no financial interest in any of the organizations mentioned above.