Autonomous weapons and stability

Scharre, Paul

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT

Unless another licence is stated on the immediately following page this work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

• Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
• Non Commercial: You may not use this work for commercial purposes.
• No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Autonomous Weapons and Stability
Paul Scharre

March 2020

Thesis Submitted in fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
in the
Department of War Studies
Faculty of Social Science & Public Policy
King’s College London
Copyright © 2019 by Paul Scharre

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.
ABSTRACT

Autonomous Weapons and Stability
Paul Scharre

Militaries around the globe are developing increasingly autonomous robotic weapons, raising the prospect of future fully autonomous weapons that could search for, select, and engage targets on their own. Autonomous weapons have been used in narrow circumstances to-date, such as defending ships and vehicles under human supervision. By examining over 50 case studies of autonomous military and cyber systems in use or currently in development around the globe, this thesis concludes that widespread use of fully autonomous weapons would be a novel development in war. Autonomous weapons raise important considerations for crisis stability, escalation control, and war termination that have largely been underexplored to-date.

Fully autonomous weapons could affect stability in a number of ways. Autonomous weapons could increase stability by replacing human decision-making with automation, linking political leaders more directly to tactical actions. Accidents with autonomous weapons could lead to the opposite, however, causing unintended lethal engagements that escalate a crisis or make it more difficult to terminate conflict. Nine total mechanisms are explored by which fully autonomous weapons could affect stability, four of which would enhance stability and five of which would undermine stability. The significance of these mechanisms is assessed, and the most prominent factor is the extent to which autonomous weapons increase or decrease human control. If autonomous weapons were likely to operate consistent with human intent in realistic military operational environments, then widespread autonomous weapons would improve crisis stability, escalation control, and war termination by increasing political leaders’ control over their military forces and reducing the risk of accidents or unauthorized escalatory actions. If, however, unanticipated action by autonomous weapons is likely in realistic environments, then their use would undermine crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could escalate conflicts or make it more difficult to end wars.

The competing frameworks of normal accident theory and high-reliability organizations make different predictions about the likelihood of accidents with complex systems.
Over 25 case studies of military and non-military complex systems in real-world environments are used to test these competing theories. These include extensive U.S. experience with the Navy Aegis and Army Patriot air and missile defense systems and their associated safety track records.

Experience with military and civilian autonomous systems suggest that three key conditions exist that will undermine reliability for fully autonomous weapons in wartime: (1) the absence of routine day-to-day experience under realistic conditions, since testing and peacetime environments will not perfectly replicate wartime conditions; (2) the presence of adversarial actors; and (3) the lack of human judgment to flexibly respond to novel situations. The thesis concludes that normal accident theory best describes the situation of fully autonomous weapons and that militaries are highly unlikely to be able to achieve high-reliability operations with fully autonomous weapons. Unanticipated lethal actions are likely to be normal consequences of using fully autonomous weapons and militaries can reduce but not entirely eliminate these risks.

Recent and potential future advances in artificial intelligence and machine learning are likely to exacerbate these risks, even as they enable more capable systems. This stems from the challenge of accurately predicting the behavior of machine learning systems in complex environments, their poor performance in response to novel situations, and their vulnerability to various forms of spoofing and manipulation.

The widespread deployment of fully autonomous weapons is therefore likely to undermine stability because of the risk of unintended lethal engagements. Four policy and regulatory options are explored to mitigate this risk. Their likelihood of success is assessed based on analysis of over 40 historical cases of attempts to control weapons and the specific features of autonomous weapons. The thesis concludes with three potentially viable regulatory approaches to mitigate risks from fully autonomous weapons.
# TABLE OF CONTENTS

Abstract ................................................................................................................................. 3

List of Tables, Images, and Illustrations ........................................................................... 7

Overview and Key Research Questions .............................................................................. 8

- Introduction ...................................................................................................................... 8
- Topic Overview ............................................................................................................... 14
- Background and Existing Literature ............................................................................. 17
- Research Questions ....................................................................................................... 41
- Methodology .................................................................................................................. 43
- Outline of the Report and Preview of Conclusions ....................................................... 55

Part I: Autonomy and Autonomous Weapons ................................................................. 57

- The AI and Robotics Revolution .................................................................................. 57
- The Nature of Autonomy ............................................................................................... 66
- A History of Automation and Autonomy in Weapons ................................................. 71

Part II: Autonomy in Weapons Today and Future Trends ............................................... 91

- Autonomy in Robotic Platforms .................................................................................... 91
- Autonomy in Missiles and Sensors ............................................................................... 101
- Commercially-Available Components for Autonomous Weapons ......................... 112
- Cyberspace and Autonomous Weapons ..................................................................... 113

Part III: Stability and Autonomous Weapons .................................................................. 126

- The Concept of Stability .............................................................................................. 127
- Autonomous Weapons and Stability .......................................................................... 132

Part IV: Controlling Autonomous Weapons: Risk, Predictability, and Hazard ............. 156

- Understanding Accident Risk: Normal Accident Theory vs. High-Reliability Organizations .................................................................................................................. 156
- Case Studies of Non-Military Autonomous Systems in Competitive and Uncontrolled Environments ........................................................................................................ 165
- Automation as a Source of Both Reliability and Failure ............................................. 200
Nuclear Weapons as a Case Study in Risk Management .............................................. 213
Summary: The Inevitability of Accidents ................................................................... 216

Part V: Learning Machines ...................................................................................... 222
  Machine Learning and Neural Networks .............................................................. 222
  Artificial General Intelligence .............................................................................. 227

Part VI: Operational Factors: The Military Value and Potential Consequences of
  Autonomous Weapons ...................................................................................... 240
  Centaur Warfighting: Human-Machine Teaming ............................................... 241
  Flash War: Autonomy and Escalation ................................................................. 253

Part VII: Risk Mitigation: Policy Options and Historical Precedents ............... 265
  Why Some Bans Work and Others Fail ............................................................. 265
  Policy Options and Prospects for Regulation ................................................... 274

Conclusion ............................................................................................................. 285

Appendices ........................................................................................................ 291
  A. Abbreviations ................................................................................................. 291
  B. Definitions of Autonomous Weapons ........................................................... 294
  C. Autonomous Weapons in Use Today: Selected Examples .......................... 298
  D. Historical Weapons Bans ............................................................................. 310

Bibliography ....................................................................................................... 316
  Interviews ........................................................................................................ 316
  Original Sources .............................................................................................. 318
  Secondary Sources ......................................................................................... 333
# LIST OF TABLES, IMAGES, AND ILLUSTRATIONS

## Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARM vs. Harpy</td>
<td>82</td>
</tr>
<tr>
<td>Potentially Stabilizing Effects of Autonomous Weapons</td>
<td>148</td>
</tr>
<tr>
<td>Potentially Destabilizing Effects of Autonomous Weapons</td>
<td>150</td>
</tr>
<tr>
<td>Comparison of Patriot vs. Aegis Missile Defense Systems</td>
<td>199</td>
</tr>
<tr>
<td>Historical Weapons Bans</td>
<td>310</td>
</tr>
</tbody>
</table>

## Images

<table>
<thead>
<tr>
<th>Image</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomahawk Anti-Ship Missile Mission Profile</td>
<td>83</td>
</tr>
<tr>
<td>Targeting image from F-15 strike in Iraq (12 kilobytes in size)</td>
<td>249</td>
</tr>
</tbody>
</table>

## Illustrations

<table>
<thead>
<tr>
<th>Illustration</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observe, Orient, Decide, Act (OODA) loop</td>
<td>63</td>
</tr>
<tr>
<td>Semi-Autonomous Operation (human in the loop)</td>
<td>67</td>
</tr>
<tr>
<td>Supervised Autonomous Operation (human on the loop)</td>
<td>68</td>
</tr>
<tr>
<td>Fully Autonomous Operation (human out of the loop)</td>
<td>68</td>
</tr>
<tr>
<td>Spectrum of Intelligence in Machines</td>
<td>70</td>
</tr>
<tr>
<td>Weapon System OODA Loop</td>
<td>77</td>
</tr>
<tr>
<td>Semi-Autonomous Weapon System (human in the loop)</td>
<td>78</td>
</tr>
<tr>
<td>Supervised Autonomous Weapon System (human on the loop)</td>
<td>79</td>
</tr>
<tr>
<td>Fully Autonomous Weapon System (human out of the loop)</td>
<td>81</td>
</tr>
<tr>
<td>Deep Neural Network</td>
<td>111</td>
</tr>
<tr>
<td>Risk of Delegating Autonomy to a Machine</td>
<td>210</td>
</tr>
<tr>
<td>Why Use Supervised Autonomy?</td>
<td>247</td>
</tr>
<tr>
<td>Why Use Full Autonomy?</td>
<td>252</td>
</tr>
<tr>
<td>Types of Weapons Bans</td>
<td>270</td>
</tr>
</tbody>
</table>
OVERVIEW AND KEY RESEARCH QUESTIONS

Introduction

The man who saved the world

On the night of September 26, 1983, the world almost ended. It was the height of the Cold War with nuclear arsenals bristling on both sides. Earlier that spring, President Reagan had announced the Strategic Defense Initiative, nicknamed “Star Wars,” a planned missile defense shield that threatened to upend the Cold War’s delicate balance of terror. The Soviet Union was on hair trigger alert. Just three weeks earlier, the Soviet military had shot down a commercial airliner flying from Alaska to Seoul that had strayed into Soviet airspace. Two hundred and sixty-nine people were killed, including a U.S. congressman.

Soviet leadership feared a surprise U.S. nuclear attack. To give them advance warning, the Soviet Union had recently deployed a satellite early warning system called Oko to watch for U.S. missile launches. Just after midnight on September 26th, the system reported it had detected one: The United States had launched a nuclear missile at the Soviet Union.

Lieutenant Colonel Stanislav Petrov was on duty that night in bunker Serpukhov-15 outside Moscow. It was his responsibility to report the missile launch up the chain of command to his superiors. In the bunker, sirens blared. A giant red backlit screen flashed “launch,” warning him of the detected missile launch.

Petrov was uncertain, however. Oko was new, and he worried that the launch might be an error, a bug in the system. He waited.

---

1 This thesis comports with King’s College London PhD regulations (R15.1 to R15.6) in “R15 Academic Regulation (Appendix) 2017/18,” https://www.kcl.ac.uk/campuslife/acservices/Academic-Regulations/assets-17-18/R15.pdf. Consistent with regulations R15.1 and R15.6, the greater proportion of investigations were conducted during the period of registration for the degree. Consistent with regulation R15.3, some of the research has already been published in Paul Scharre, Army of None: Autonomous Weapons and the Future of War (New York, NY: W. W. Norton & Co., 2018), copyright © 2018 by Paul Scharre, used by permission of W. W. Norton & Company, Inc. All of the work in this thesis is the author’s own. Consistent with regulation R15.6(k), ethical approval to conduct interviews was granted by the King’s College London Research Ethics Office.


Another launch. Two missiles were inbound. Then another. And another. And another.  

Five missiles launched from the United States towards the Soviet Union. The flashing screen switched from “launch” to “missile strike.” The system reported the highest confidence level. There was no ambiguity. A nuclear strike was on its way.

Protocol was clear – the military officer on duty was required to report the launch up the chain of command. Soviet military command would have only minutes to decide what to do before the missiles would explode over Moscow.

But Petrov had a funny feeling in his gut. Why would the United States launch only five missiles? It didn’t make sense. A real surprise attack should be massive, an overwhelming strike to wipe out Soviet missiles on the ground. Petrov wasn’t convinced the attack was real. But he wasn’t certain it was a false alarm, either.

He had only minutes to decide. With one eye on the computer readouts, Petrov called the ground-based radar operators for confirmation. If the missiles were real, as they arced over the horizon they should come up on Soviet ground-based radars as well. But the ground radars saw no missiles.

Petrov felt like he was sitting on a hot frying pan. He judged the odds of the strike being real were 50-50, a coin flip. But the consequences were enormous. Make this decision wrong and millions would die. He didn’t want to make a mistake. He needed more information. He needed more time. He wasn’t going to get either.

His job was clear. His duty was to report the launch up the chain of command. All he had to do was pick up the phone. But in the heightened climate of fear, Petrov feared how Soviet leadership might react. They could start World War III.

Petrov went with his gut. He called his superiors to inform them the system was malfunctioning. The strike was false, he told them.

Petrov was right. Sunlight reflecting off cloud tops had triggered a false alarm in Soviet satellites. There was no attack. The system was wrong. Humanity was saved from potential Armageddon by a human “in the loop.”

What would a machine have done? Whatever it was programmed to do.

---

4 Ibid.

5 Hoffman, “I Had a Funny Feeling in My Gut.”

6 Aksenov, “Stanislav Petrov: The Man Who May have Saved the World.”
Technology has brought us to the brink of a crucial threshold in humanity’s relationship with war. In future wars, machines could make life and death engagement decisions all on their own. Militaries around the globe are racing to deploy robots in the air, on the ground, and at sea. They are increasingly autonomous and many are armed. Over 90 countries have drones patrolling the skies. Drones operate under human control, for now, but what happens when a Predator drone has as much autonomy as a self-driving car? What authority should militaries be willing to delegate to machines when the ultimate decision is at stake – life and death?

Over 30 nations already have defensive autonomous weapons for situations in which the speed of engagements is too fast for humans to respond. They are used to defend ships and bases against saturation attacks from rockets and missiles. These systems are supervised by humans who can intervene if necessary, but more autonomy is creeping into a wide range of weapons. One such weapon, the Israeli Harpy drone, already crosses the line to a fully autonomous weapon. Unlike the Predator drone, which is controlled by a human, the Harpy can search a wide area for enemy radars and, once it finds one, kamikaze into it without asking permission. It’s been sold to a handful of countries – Turkey, China, India, and South Korea – and China has reverse-engineered their own variant. Wider proliferation is a definite possibility, and the Harpy may only be the beginning.

Militaries around the globe are deploying armed robots with increasing autonomy. South Korea has deployed a robotic sentry gun to the demilitarized zone bordering North Korea. Israel has used armed ground robots to patrol its Gaza border. Russia is building a suite of armed ground robots for war on the plains of Europe. Over twenty nations have or are in the process of acquiring armed drones.7

These developments are facets of a deeper technology trend that is sweeping through human society: the rise of artificial intelligence (AI). Some have called AI the “next industrial revolution.”8 Technology guru Kevin Kelly has compared AI to electricity. Just as electricity infuses objects all around us, bringing them to life with

---


power, so too will AI infuse everyday objects, bringing them to life with intelligence. In physical systems, AI enables more sophisticated and autonomous robots, from warehouse robots to next-generation drones. In non-physical systems, AI can help process large amounts of data and make decisions for a variety of tasks, making medical diagnoses, programming subway repair schedules, and powering Twitter bots. In warfare, AI systems can be used to process information to help humans make decisions – or they can be delegated authority to make decisions on their own.

The rise of artificial intelligence will transform warfare. In the early twentieth century, militaries harnessed the industrial revolution to bring tanks, aircraft, and machine guns to war, unleashing destruction on an unprecedented scale. The industrial revolution enabled the creation of machines that were physically stronger and faster than humans, at least for tailored tasks. Similarly, the AI revolution is enabling the creation of machines that are smarter and cognitively faster than humans for narrow tasks. Many military applications of AI are uncontroversial – improved logistics, cyber defenses, decision aids to assist humans, or robots for medical evacuation, resupply, or surveillance. The infusion of AI into weapons raises difficult questions, however. While automation is already used for a variety of functions in weapons today, in most cases it is still humans choosing the targets and pulling the trigger. Whether that will continue is unclear. Most countries have kept silent on their plans, but a few have signaled their intention to not hold back on autonomy. Senior Russian military commanders have said they envision that in the near future a “fully robotized unit will be created, capable of independently conducting military operations.” U.S. Defense Department officials have stated fully autonomous weapons should be “on the table.”

Better than human?

Armed robots deciding on their own whom to kill might sound like a dystopian sci-fi nightmare, but some argue autonomous weapons might make war more humane. The same kind of automation that could allow self-driving cars to avoid pedestrians

---


could also be used to avoid civilian casualties in war. Unlike human soldiers, machines would never get angry or seek revenge. They would never fatigue or tire. Airplane autopilots have dramatically improved safety for commercial airliners, saving countless lives. Could autonomy do the same for war? Countless tasks once viewed as impossible for machines have now fallen to the seemingly endless march of machines surpassing human skills: chess, Jeopardy, Go, flying, driving, and stock trading. Is war next?

New AI techniques such as deep learning neural networks have shown startling advances in visual object recognition, facial recognition, and sensing human emotions. It isn’t hard to imagine future sensors that could outperform humans in discriminating between a person holding a rifle and one with a rake. Yet computers still fall far short of humans in understanding context and interpreting meaning. AI programs today can identify objects in images, but they can’t place these objects into a broader context to understand what’s happening.

Some decisions in war are straightforward. Sometimes the enemy is easily identified and the shot is clean. Some decisions however, like the one Stanislav Petrov faced, require understanding the broader context. Sometimes doing the right thing entails breaking the rules.

The debate

The prospect of fully autonomous weapons that could make life or death decisions on their own has sparked a vibrant debate, with some calling for a pre-emptive ban on autonomous weapons. Over 3,000 robotics and artificial intelligence experts have called for a ban on offensive autonomous weapons. They are joined by over 60 non-governmental organizations who are part of a Campaign to Stop Killer Robots. Science and technology luminaries such as Stephen Hawking, Elon Musk, and Apple co-founder Steve Wozniak have spoken out against autonomous weapons, warning they could spark a “global AI arms race.” The International Committee of the Red Cross has similarly expressed concerns about autonomous weapons, although they have yet to call for a ban, while the UN special rapporteur for extrajudicial killings has called for a moratorium on autonomous weapon development. Discussions among states on what to

---


do about autonomous weapons are underway. Since 2014, nations have gathered annually in Geneva, Switzerland as part of the United Nations (UN) Convention on Certain Conventional Weapons (CCW) to discuss lethal autonomous weapons.

Much of the discussion on autonomous weapons to-date has focused on legal, moral, and ethical issues, but autonomous weapons also raise important concerns for stability. By removing the human “from the loop” for decisions to use force against specific targets, autonomous weapons would remove a potentially valuable failsafe against accidents or unintended actions on the battlefield. Could autonomous weapons lead to a condition of “fragile stability,” where crises could quickly escalate out of control? Automated stock trading has led to “flash crashes” on Wall Street. Could the interaction between autonomous weapons on the battlefield lead to “flash wars,” unintended accidents with catastrophic outcomes?

Today’s “narrow” AI systems are brittle. They can often exceed human performance in narrow tasks, such as driving, but when pushed beyond the boundaries of their programming, they can fail, and fail badly. Current narrow AI systems lack the flexibility of human general intelligence to adapt to novel circumstances. Could future more advanced AI systems be built smart enough to understand the broader context as Stanislav Petrov did, overcoming these limitations? If so, what new challenges would more advanced AI systems with human-like intelligence bring?

Finally, if autonomous weapons do raise concerns for stability, what steps can nations take to address them, unilaterally and collectively? Humanity’s track record for controlling dangerous technology is mixed. Attempts to ban weapons that were seen as illegitimate date back to antiquity. Many of these attempts have failed, including early 20th century attempts to ban submarines and airplanes. Even those that have succeeded, such as the ban on chemical weapons, rarely stop rogue regimes such as Bashar al-Assad’s Syria or Saddam Hussein’s Iraq. However, during the Cold War, the U.S. and U.S.S.R. cooperated to avoid weapons or deployment postures that were seen as destabilizing, such as anti-ballistic missile weapons, intermediate range nuclear-capable missiles, or placing nuclear weapons in space or on the seabed.

Autonomous weapons that could make life or death decisions on their own raise difficult legal or ethical issues, but also raise practical safety concerns about controllability and risk. Autonomous weapons slipping beyond human control could have grave consequences for crisis stability, escalation management, and war termination. A “flash war” would benefit no one. Stability is an important consideration that should be part of the dialogue about the consequences of autonomous weapons.
Topic Overview

Nations around the globe are deploying military robots increasing in autonomy. Automation and autonomy is also increasing in many non-robotic systems, including munitions, air and missile defenses, and cyber systems. These developments have sparked a growing literature on autonomous weapons and the appropriate role of human control in the use of force. However, the current literature suffers from two significant shortcomings. First, much of the literature is advocacy-based by activists campaigning for a ban on “killer robots”, although there is a growing body of academic literature on autonomous weapons. Second, much of the literature on autonomous weapons engages the topic from a legal, ethical, or moral perspective. While these are important considerations, autonomous weapons also raise important concerns for stability. Delegating the use of force to an autonomous system could create novel challenges for stability, including escalation management in crises and war termination.

This thesis examines the implications of autonomous weapons for stability. The central research question is:

Do autonomous weapons pose challenges to stability and, if so, what are feasible risk mitigation measures that nations could take?

The first part of this question will be examined by an exploration of the nature of autonomous weapons and their intersection with the concept of stability, including crisis stability, escalation dynamics, and war termination. One of the central concerns with autonomous weapons is the risk that comes from delegating lethal decision-making to a machine. Simply put, the risk – both the probability and consequence – that the weapon system attacks something that the human who put it into operation did not want it to attack. The Department of Defense refers these situations as “unintended engagements,” a concept that also includes excessive collateral damage on otherwise appropriate targets. This could occur for a variety of reasons: simple mechanical malfunctions, software glitches, system failures stemming from the interaction of different components, inadequate testing, faulty design, insufficient training, poor human-machine interfaces, human error, unanticipated interactions with the environment, and adversarial hacking, spoofing, or behavioral manipulation.

While some of these concerns are exacerbated by system complexity and/or autonomy, in general these risks are not unique to autonomous weapons. Fratricide, civilian casualties, and accidents are unfortunate realities of war. Autonomous weapons pose a unique challenge, however. A failure with an autonomous weapon could result in a “runaway gun” that attacks inappropriate targets on its own for some length of time before a human is able to reassert control. In the worst case, a weapon could continue engaging inappropriate targets until it exhausts its magazine. Depending on the specific weapon, target type, magazine capacity, and context for use, such an accident could lead to significant harm.

In addition to attacks on friendly forces or civilians, unintended engagements could also include attacks on enemy forces that would otherwise be legitimate targets, but in a time, place, or manner other than what the human operator intended. During periods of brinkmanship, crises, or limited wars, these engagements could lead to unintended escalation, undermining stability. Because of the speed of automation, these engagements could happen quite quickly, causing crises to rapidly spiral out of control.

On the other hand, some forms of autonomy could increase stability. Automated defensive systems could help buy time for humans to make decisions in a crisis, and automated information systems could help humans quickly process large amounts of confusing or conflicting information. Automatic response systems could increase deterrence by effectively tying a nation’s hands and reducing uncertainty, if such automaticity could be convincingly conveyed to an adversary. If autonomous weapons were viewed as unpredictable, on the other hand, uncertainty about their risks could potentially induce caution in edging down the slippery slope to war – the “madman” theory of stability applied to autonomous machines.

Therefore, a critically important dimension to the effect of autonomous weapons on stability is the extent to which their operation is controllable and predictable. This includes both the behavior of the system itself as well as its interaction with adversary systems in competitive environments. Because fully autonomous weapons generally do not yet exist, with some rare exceptions, this thesis examines over 80 total case studies of related systems. These include existing weapon systems with varying degrees of autonomy as well as non-military autonomous systems operating in competitive and/or unconstrained environments. These case studies are examined in order to understand the extent to which human users can accurately anticipate and control the behavior of complex, highly autonomous systems in unconstrained or adversarial environments. These include:
• Accidents involving complex automated or autonomous systems in non-military settings, including the F-22 international dateline incident and Air France 447 crash.

• Accidents involving automation in the competitive, high-speed adversarial environment of stock trading including the causes of the May 6, 2010 flash crash and subsequent risk mitigation measures.

• The 2003 Patriot fratricides, in which the highly automated U.S. Patriot air defense system shot down two friendly aircraft during the initial phases of the Iraq war;

• The performance of the Aegis combat system, a ship-based defensive human-supervised autonomous weapon system operated by the U.S. Navy.

• The behavior of cutting-edge present-day machine learning systems.

These case studies are examined through the lens of two theories regarding risk in complex systems: normal accident theory and high-reliability organizations, which have competing predictions regarding the risk of accidents in complex systems. Normal accident theory predicts that prudent measures can be taken to minimize risk, but that accidents are inevitable in complex, tightly-coupled systems. An alternative theory is that, with certain procedures, high-reliability organizations can achieve very low accidents rates in high-risk settings. This thesis evaluates the applicability of these theories to autonomous weapons, and the extent to which case studies suggest autonomous weapons would be controllable and predictable to users. In addition, the safety record of nuclear weapons is examined as one measure of the capacity of militaries to safely manage hazardous weapon systems.

At the foundation of this problem of controllability is the brittle nature of machine intelligence. Unlike human intelligence, which is broad and allows humans to flexibly adapt to a range of circumstances, machine intelligence – at least given today’s technology – is quite narrow. AI systems can often outperform humans in narrow tasks, such as driving cars or playing Go, but cannot adapt to novel circumstances. If pushed outside the bounds of their tasks, they will fail, and often fail badly. Some researchers have speculated that future machines could be built which would exhibit human-like
artificial general intelligence (AGI). In principle, a machine that exhibited general intelligence would address many of the concerns about unintended engagements. A true AGI could, as humans do, reason about what commander’s intended, not merely what their orders said, and adjust their actions accordingly. The prospects for developing AGI are unknown and estimates for when it might developed are wildly speculative. Nevertheless, this study will briefly examine the nascent (and admittedly extremely speculative) literature on AGI and consider AGI’s implications for autonomous weapons and stability.

Finally, if autonomous weapons pose risks for stability, this thesis considers potential risk mitigation measures that nations might undertake, both unilaterally and cooperatively. There is a rich history of international arms control and unilateral self-regulation, dating back millennia. The thesis examines examples in which efforts to restrict weapons have been successful and unsuccessful and draws conclusions about the prospects for managing the risks of autonomous weapons.

This thesis benefits several audiences: First, it contributes to the academic literature on strategic stability by applying existing theories to an emerging technology. Second, it helps inform real-world international discussions on autonomous weapons. Militaries are grappling today with the question of how much autonomy and human control to design into future weapon systems. Additionally, states, international organizations, and non-governmental organizations are engaged in ongoing discussions about what, if any, international norms, treaties, regimes, or regulations would be useful with regard to autonomous weapons. While not policy prescriptive in nature, this thesis nevertheless generates policy-relevant findings to help inform those debating increasing autonomy in weapons.

**Background and Existing Literature**

Stability is an important consideration for emerging weapons technologies and has a rich intellectual history. Stability\(^\text{16}\) encompasses a number of different concepts, including first strike stability, escalation control, and arms race stability, which have varying degrees of relevance to autonomous weapons. In addition, autonomous weapons may pose novel challenges for war termination, an important consideration sometimes included within the broad umbrella of stability.

---

\(^{16}\) “Strategic stability” is sometimes used to refer specifically to nuclear weapons—hence, the use of “stability” here when used in reference to autonomous weapons.
The concept of stability emerged during the 1950s among U.S. nuclear theorists attempting to grapple with the implications of this new and powerful weapon. U.S. officials began to worry as early as 1947 that the sheer scale of nuclear weapons’ destructiveness gave an advantage to whichever nation struck first, potentially leading the Soviet Union to launch a surprise nuclear attack. This vulnerability of U.S. nuclear forces to a surprise Soviet attack therefore gave the United States an incentive to themselves strike first, if war appeared imminent. Knowing this, of course, only further incentivized the Soviet Union to strike first in the event of possible hostilities. This dangerous dynamic captures the essence of first strike instability, a situation in which adversaries face off like gunslingers in the wild west, each poised to strike first at the slightest hint that the other might reach for his gun. As Thomas Schelling explained the dilemma, “we have to worry about his striking us to keep us from striking him to keep him from striking us.” The danger is that instability itself can create a self-fulfilling prophecy where one side attacks first, fearing an attack from the other side.

While U.S. strategists considered immediate steps to reduce the vulnerability of U.S. nuclear forces to a surprise attack, over time the concept evolved of a desire for “stability” more generally. The concept of stability is broader than merely protecting one’s own forces from a surprise attack; it takes into account the perspective of both sides in a conflict. Stability also includes avoiding deploying one’s military forces in such a way that an adversary would see himself to be at risk of a surprise attack, thus incentivizing him to strike first. A stable situation, Thomas Schelling described, is “when neither in striking first can destroy the other’s ability to strike back,” thus avoiding the unstable gunslinger dynamic that incentivizes each side to strike first, if he believes the other might attack.

The concept of strategic stability is a direct analogy from the physical world. A stable equilibrium is one that, if disturbed by an outside force, returns to its original equilibrium. A ball sitting at the bottom of a bowl is at a stable equilibrium. If the ball is moved slightly, it will return to the bottom of the bowl. Conversely, an unstable equilibrium is one where a slight disturbance will cause the system to rapidly transition

---


19 Ibid.
to an alternate state. A pencil balanced on its tip is at an unstable equilibrium. Any slight disturbance will cause the pencil to tip over to one side. When nuclear war is on the line, a condition analogous to a pencil balanced on its tip is a precarious situation to be in.

Over time, several different variants of stability emerged. The original concept of reducing incentives for a surprise attack became known as “first strike stability.” First strike stability is a condition where competitors do not have an incentive to strike first. A related concept is one of “crisis stability.” Crisis stability is related to first-strike stability but takes a somewhat broader view and is concerned with avoiding any unwanted conditions that might escalate a crisis. These could include perverse incentives for deliberate escalation, such as the vulnerabilities that can cause first-strike instability. Crisis stability could also be undermined by accidental escalation, for example from lower-level commanders taking matters into their own hands; automatic escalation, such as by pre-delegating authority for responses that, in practice, policymakers might not want to execute; or miscalculation of an adversary’s actions or intentions that leads to escalation. Finally, the concept of stability can be extended in time to the notion of arms race stability, a desirable condition where countries might avert a costly and wasteful arms race because it would not gain them any advantage over an adversary, thus creating a stable condition.20

Stability has proven to be a vitally important intellectual tool for mitigating potential risks from nuclear weapons. Throughout the Cold War, the U.S. and U.S.S.R engaged in a number of unilateral and cooperative measures designed to increase stability and avoid potentially unstable situations that might propel the nations to war in a crisis. While it is impossible to definitely prove that these measures had the desired effect – after all, a counterfactual experiment would consist of deliberately attempting to incite nuclear war – the record of near-miss nuclear incidents throughout the Cold War and afterward suggests, if anything, these stability-inducing measures could be even more robust.

As new military technologies emerge, stability is an important lens through which they should be evaluated. Many technologies will not significantly affect stability one way or the other, but some military technologies do have strategic effects. Increased

competition in space and cyberspace, for example, are two areas with strategic effects.\textsuperscript{21} The effect of this competition on stability should be continually evaluated. Similarly, autonomous weapons that could select and attack targets on their own should be evaluated for their effect on stability.

*What is an autonomous weapon?*

Automation has long been used in weapon systems in various forms. For over seventy years, militaries have employed “fire and forget” homing munitions that possess onboard seekers to home in on enemy targets. Increasing autonomy is enabling future systems, however, that could search for targets over wide areas and destroy them on their own. These autonomous weapon systems would be weapons that, once activated, would be intended to select and engage targets where a human had not decided those specific targets are to be engaged.

At least 30 nations already possess human-supervised autonomous weapon systems that are used to defend vehicles and bases from short-warning saturation attacks.\textsuperscript{22} These include air and missile defense systems; counter-rocket, artillery, and mortar systems; and active protection systems to protect ground vehicles from incoming projectiles. Human operators supervise their operation and can intervene to halt their functioning if necessary, although not necessarily before an incident occurs. (Human operators may, in fact, be unaware that the system is performing inappropriately until an accident occurs.) Human operators also generally have physical access with the ability, at least in principle, to manually disable the system in case it fails to respond to software commands.

These systems are essential for militaries to effectively defend bases and vehicles in an era of increasingly sophisticated precision-guided weapons. Without autonomous modes of operation, human operators could be overwhelmed by short-warning saturation attacks that could outstrip their ability to respond to each incoming


threat. These systems are defensive in nature, likely contributing to stability, and in some cases are used to defend civilian populations.

Even with human supervision, these systems are not flawless, however. In 2003, the U.S. Patriot air defense system was engaged in two fratricide incidents where it shot down friendly aircraft, killing the pilots. The accidents resulted from a complex mix of inadequate training, poor system design, opaque human-machine interfaces, unexpected conditions on the battlefield, and an operator culture that placed excessive trust in automation.\textsuperscript{23} The fact that a human was “in the loop” for both incidents did not prevent them from occurring. These incidents point to some of the increased risks of highly automated weapons when employed in real-world operating conditions.

Increasing autonomy continues to be incorporated into future military weapon systems, not just drones and uninhabited vehicles but also next-generation munitions. The anti-radar Israeli Harpy is one example.\textsuperscript{24} Additionally, some proponents of a ban have raised questions about a number of next-generation missiles with advanced targeting features, including the UK Brimstone, Norway’s Joint Strike Missile (JSM), and the U.S. Long-Range Anti-Ship Missile (LRASM).\textsuperscript{25} Publicly posted data on their functionality is often ambiguous as to their degree of autonomy in selecting and engaging targets.\textsuperscript{26} This highlights one of the difficulties in arms control with


\textsuperscript{24} Israel Aerospace Industries, “Harpy Loitering Weapon,” \texttt{www.iai.co.il/2013/16143-16153-en/IAI.aspx}. Israel has sold the Harpy to Turkey, South Korea, India, and China. See “Harpy Air Defense Suppression System,” Defense Update, updated March 4, 2006, \texttt{defense-update.com/directory/harpy.htm}. China is alleged to have reverse-engineered its own variant. See “UAV/UCAV,” Chinese Military Aviation, May 7, 2014, \texttt{chinese-military-aviation.blogspot.ch/p/uav.html}. The Harpy is frequently referred to as an unmanned aerial vehicle (UAV) or unmanned combat air vehicle (UCAV), but the Harpy is essentially a wide-area loitering cruise missile.


autonomous weapons. Because the difference between an autonomous weapon and a semi-autonomous one may be in the software, transparency and verification may be difficult.

These next-generation missiles also demonstrate one of the challenges with incrementally advancing autonomy over time. Even militaries building autonomous weapons may not see a bright line between the weapons of today and future autonomous ones. Militaries may incrementally add more autonomous functionality without necessarily having a considered debate or even a conscious decision to pursue autonomous weapons, cognizant of the strategic risks and benefits.

Why would militaries want autonomous weapons?

While no nation has stated that they are pursuing autonomous weapons, few have ruled them out either. Militaries have many motivations to pursue increasingly autonomous features in military systems, including greater precision, reliability, reduced personnel costs, and the ability to conduct operations with uninhabited and robotic systems in communications-degraded or -denied environments.

In many cases, the advantages of increased autonomy could be gained while still retaining a human “in the loop” to authorize specific targets for engagement. On-board targeting algorithms that could cue potential targets to human operators could reduce bandwidth requirements dramatically down to only hundreds of kilobits per second, allowing operations in communications-degraded environments. Nevertheless, retaining human authorization for specific targets would still require some bandwidth as well as some time for human operators to make an informed decision – at least several seconds – depending on the target and the complexity of the environment.

Militaries might wish to delegate engagement decision authority to an autonomous system if acquiring human authorization is not feasible. This could be the

27 As of November 2018, twenty-seven nations explicitly supported a ban on autonomous weapons: Pakistan, Ecuador, Egypt, the Holy See, Cuba, Ghana, Bolivia, Palestine, Zimbabwe, Algeria, Costa Rica, Mexico, Chile, Nicaragua, Panama, Peru, Argentina, Venezuela, Guatemala, Brazil, Iraq, Uganda, Austria, Colombia, Djibouti, El Salvador, and Morocco (in order of supporting a ban). Campaign to Stop Killer Robots, “Country Views on Killer Robots,” November 22, 2018, https://www.stopkillerrobots.org/wp-content/uploads/2018/11/KRC_CountryViews22Nov2018.pdf. Some major military powers, such as the United States and Russia, have explicitly left the door open to autonomous weapon development in the future, although stating that they were not developing them at this time.

case if communications with uninhabited combat vehicles are degraded or denied. This could also occur if engagement timelines are so compressed that keeping a human in the loop is no longer feasible. This is the rationale for why nations have defensive human-supervised autonomous weapons today and a similar logic could apply in offensive roles as well.

Militaries might also desire autonomous weapons simply out of a fear that others might be developing them and that they could yield an important battlefield advantage, even if their specific application was not yet known.

Finally, the incremental nature of autonomous development in many areas, such as automobiles, points to the potential for militaries to creep into autonomous weapons, perhaps without even making a deliberate strategic decision to pursue them.

Legal and ethical aspects of autonomous weapons have been extensively explored

Autonomous weapons raise a number of concerns, including legal, moral, and ethical concerns regarding human accountability and responsibility for the use of force. These issues have been engaged in a growing literature on autonomous weapons. While much of the literature on autonomous weapons comes from a position of advocacy, there is nevertheless an extensive debate on the legal and ethical issues surrounding autonomous weapons. Matthew Waxman, Ken Andersen, Daniel Reisner, and Rebecca Crootof have written on many of the legal issues surrounding autonomous weapons.29 A number of scholars such as Peter Asaro, Robert Sparrow, and B.J. Strawser have examined the moral and ethical dimensions of autonomous weapons.30 Some of the


ethical concerns surrounding human dignity and autonomous weapons have been most elegantly argued by Christof Heyns, the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions.\textsuperscript{31}

Safety and controllability

Autonomous weapons also raise important considerations for safety and controllability. The essence of autonomy is delegating control of a task to a machine. Yet, autonomous systems are still designed and employed by humans for a purpose. How will human operators retain meaningful and effective control over weapons with increasing autonomy, particularly if autonomous weapons might be most useful in situations where there are no communications or the speed of operations outpaces human reaction times? And how can militaries ensure failsafe operations in complex real-world environments, subject to adversary hacking, spoofing, physical attacks, and deception? A loss of control could significantly challenge policymakers’ ability to manage escalation in a crisis, potentially even leading to accidental or inadvertent escalation. Safety and controllability issues surrounding autonomous weapons are rarely addressed, and when they are, they are done so generally within two, fairly limited, contexts.

The first arena in which controllability has been examined is within the framework of “meaningful human control,” a nebulous term that has been put forward by ban advocates as a positive vision for the role of autonomy in future weapons.\textsuperscript{32}


Despite a lack of a clear definition on what meaningful human control entails, the term has caught on among many concerned about autonomous weapons. Meaningful human control has been the topic of significant discussion at the UN CCW meetings on autonomous weapons. Meaningful human control was also the subject of a report by the non-advocacy UN research group, the United Nations Institute for Disarmament Research (UNIDIR). Discussion of meaningful human control generally has focused on the notion of what makes human control “meaningful” in an ethical or legal sense, focusing on human responsibility and accountability. The focus of the dialogue on meaningful human control has often been from the perspective of the psychology of the human operator and whether the operator is cognitively engaged in decision-making, or merely ceding decision-making authority to the automation. While there is an existing literature on automation bias and human-automation interaction, most of the discussion regarding meaningful human control centers around defining the problem and desired goals.

The second, critical dimension of meaningful human control – safety and controllability – pertains to the technical aspects of designing autonomous weapons that will function as intended under real-world operating conditions. Less has been written on this front, although some official U.S. Department of Defense publications have engaged on these issues.

The official U.S. Department of Defense policy on autonomy in weapons, DoD Directive 3000.09, Autonomy in Weapon Systems, engages extensively with questions of testing and safety to ensure that weapon systems that incorporate autonomy perform


34 For more on automation bias and human-automation interaction issues, see Hawley, “Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons.”
as desired under real-world operating conditions. The policy touches little on legal or ethical issues, merely stating that autonomous weapons, like all weapons, must pass a legal review and that persons who employ autonomous weapons must do so with “appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable rules of engagement (ROE).” Rather, the purpose of the DoD document is to “[establish] guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements.” Accordingly, the majority of the 15-page document covers standards for test and evaluation, software verification, training for human operators, system design, and other safety-related issues.


These documents examine safety and controllability aspects of autonomy narrowly, however, considering only the technical steps one might take to improve safety. They consider safety and controllability analogous to how one might examine the immediate safety concerns surrounding other weapons or dangerous tools, such as a hand grenade or chainsaw. They fail to even acknowledge that issues of controllability for autonomous weapons may raise broader strategic concerns – that the deployment of autonomous weapons on a large scale could, in fact, be quite dangerous and harmful to escalation management, war limitation, or war termination.

36 Ibid, 3, 7-8.
37 Ibid, 1.
Autonomous weapons and stability

While the stability dimensions of autonomous weapons concerns are not currently reflected in official government documents on autonomy, a few scholars have raised these concerns. Michael Carl Haas has written on the stability implications of autonomous weapons, raising questions regarding their implications for arms race stability, crisis stability, and escalation control. With regard to crisis stability, Haas argues:

[S]tates [who employ autonomous weapons] would be introducing into the crisis equation an element that is beyond their immediate control, but that nonetheless interacts with the human opponent’s strategic psychology. In effect, the artificial intelligence (AI) that governs the behavior of autonomous systems during their operational employment would become an additional actor participating in the crisis, though one who is tightly constrained by a set of algorithms and mission objectives.

This introduces a complex set of variables into the equation: the AI itself, its ability to accurately follow its human commander’s guidance, the opponent’s expectations of the AI, and any vulnerabilities the opponent may exploit in the AI. Exploiting vulnerabilities could take the form of direct hacking via malware or behavioral hacking by exploiting predictable AI behavior. Furthermore, communications jamming introduces another variable into the equation. Communications jamming may restrict or entirely eliminate communications with human operators, forcing robotic vehicles to operate fully autonomously. This then creates another set of reactions, perceptions, and counter-reactions surrounding the rules of engagement delegated to the AI.


42 Ibid.

Haas notes a problematic dimension of autonomous weapons is that their employment could complicate escalation management if they were not recallable.

Recallability and loss of control, clearly, are major concerns. While strike systems along the lines of the X-47B UCAS [unmanned combat air system] could initially be employed under close human supervision, it is difficult to see how they could realize their full potential in those scenarios where they offer by far the greatest value added: intelligence, surveillance and reconnaissance (ISR) and strike missions deep inside well-defended territory, where communications will likely be degraded and the electronic emissions produced by keeping a human constantly in the loop could be a dead giveaway.\textsuperscript{44}

Advanced militaries have methods of communicating in contested environments, although these are limited in bandwidth and range and may require human-inhabited (manned) aircraft nearby. Haas’ point remains valid, however. If communications with human controllers are not possible, the uninhabited vehicle will either have to be limited in its mission or, if it operates fully autonomously, then the humans employing it will have to accept the risk of a loss of control and recallability. Haas rightly points out that militaries may employ autonomous weapons in the early stages of conflict in “anti-access / area denial” zones deep inside enemy territory, where they may have to operate without communications with human controllers for extended periods of time.

During these stints inside the defended zone, [autonomous weapons] might not be fully recallable or reprogrammable, even if the political situation changes, which presents a risk of undesirable escalation and could undermine political initiatives.\textsuperscript{45}

Even if political circumstances do not change, adversaries could exploit vulnerabilities in the autonomous weapon’s targeting rules of engagement.

This could include, \textit{inter alia}, relocating important assets to busy urban settings or next to inadmissible targets, such as hydroelectric dams or nuclear-power

\textsuperscript{44} Ibid.

\textsuperscript{45} Ibid.
stations; altering the appearance of weapons and installations to simulate illegitimate targets, and perhaps even the alteration of illegitimate targets to simulate legitimate ones; large-scale use of dummies and obscurants, and the full panoply of electronic deception measures.46

All of this could lead to inadvertent escalation, because of a lack of recallability, a loss of effective control over the autonomous weapon, or simply because adversaries exploit weaknesses in the autonomous weapon’s targeting algorithms. Haas concludes that autonomous weapons raise a number of challenging issues for crisis stability: “there are scenarios in which the introduction of autonomous strike systems could result in temporary loss of high-level control over operations, and unwanted escalation (conventional or nuclear).” As a result, policymakers “should exercise prudence and caution in weighing the implications of autonomous weapons for military stability against the potential benefits of introducing these systems into an equation that is highly complex as it stands.” Haas argues that “mutual restraint” should be explored as a serious option. Haas’ concerns are particularly noteworthy because they reflect an appreciation of strategic risks associated with autonomous weapons that are not reflected in the policy documents or public statements of any government – the United States or others.

Kenneth Payne has raised similar concerns about AI more broadly, that their introduction into warfare will change strategy in profound ways. Risks could arise from AI systems that are misaligned with human intention or are unable to flexibly adapt to changing circumstances. He explains:

If the AI miscalculates what its human principals want, there is tremendous potential for catastrophic and unwarranted violence. … An AI that escalates to meet pre-specified criteria of reputation and credibility may not be able to reflect on the consequences of its actions in time to allow a change of course.47

This could lead to “unintended escalation.”48 Payne also argues that AI systems will lack the ability to empathize with human opponents, which can be valuable in deescalating conflict. Citing the Cuban Missile Crisis, he states that “in a tense stand-

46 Ibid.


off, social intelligence and theory of mind were decisive.” While Payne’s arguments are about the limitations of military AI systems in general, they certainly apply to autonomous weapons as well.

Alexander Velez-Green argues that autonomous weapons could increase or decrease stability, depending on the type of weapon and the circumstances of its deployment. As an example, Velez-Green cites the Samsung SGR-A1 robotic machine gun deployed to the South Korean demilitarized zone (DMZ), which in some reports is alleged to have a fully autonomous capability. Velez-Green argues that a fully autonomous SGR-A1 that could engage advancing North Korean troops even if communications with its human controllers was cut off would enhance deterrence, thus improving stability in a crisis. This is because the SGR-A1 is an entirely defensive weapon, a static gun that is fixed in place at the border. Velez-Green acknowledges the challenge in differentiating between offensive and defensive weapons, and concludes:

[S]tates should participate in an international dialogue to clearly define the proper uses of autonomous weapons, agree upon methods to verify that states are using these weapons properly, and designate consequences for violators.

Elsewhere, Velez-Green adopts a more nuanced position on the implications of autonomous weapons for stability. Building on concerns raised by Paul Scharre that autonomous weapons could interact in unpredictable ways on the battlefield, Velez-Green cites a number of examples of autonomous systems other than weapons behaving in surprising ways. These include the bidding war between two automated seller-bots on Amazon that led to a textbook being priced at $23.6 million; the 1983 nuclear incident sparked by a false alarm from the Soviet Oko missile early-warning system; and the 2010 stock market “flash crash” on Wall Street. Velez-Green warns that autonomous weapons interacting in unpredictable ways could start a “flash war,” and outlines a frightening scenario involving automated cyber weapons as to how it might happen.

49 Payne, “AI, warbot.”


51 Ibid.
John Borrie takes an even more pessimistic view of autonomous weapons. Drawing a parallel to “normal accident theory,” Borrie argues:

[T]here is a safety dimension to [autonomous weapons] that’s independent of whatever we think about the ethics or legality of autonomized targeting or attacks. My observation is that it may not be possible to effectively control or predict what will happen as the result of even initially simple failures in autonomized weapon systems. That’s not because of deliberate misuse or because “the robots are taking over”, but simply because it’s hard to diagnose and respond effectively to failures as quickly as is necessary to prevent catastrophic failure in complex, tightly coupled systems.

In coming to this conclusion, Borrie draws upon the history of complex, tightly coupled systems in other high-risk industries including nuclear reactors, nuclear weapons, commercial airlines, and oil drilling. He cites examples such as the Air France Flight 447, Three Mile Island, and Deepwater Horizon accidents.

In April 2016, Borrie expounded on these concerns in a talk on autonomous weapons and risk on the sidelines of the UN meetings on autonomous weapons in Geneva, Switzerland. Borrie focused on “unintentional risks” from autonomous weapons that could come if “systems don’t behave in ways intended by their designers and operators.” Drawing on the two competing theories for understanding risk in complex systems, normal accident theory and high-reliability organization theory, Borrie concluded that autonomous weapons would “fit the character of complex, tightly-coupled systems quite closely in many cases.” Similar to other analysts, Borrie drew parallels with accidents in complex tightly coupled systems in other fields, such as space travel, stock market flash crashes, and nuclear accidents. The consequences of an accident with autonomous weapons, Borrie suggested, could include “mass lethality accidents and impacts on strategic stability.”

Paul Scharre and Michael Horowitz raise similar concerns as Haas, Velez-


53 Ibid.

Green, and Borrie about the predictability or controllability of autonomous weapons.\textsuperscript{55} Scharre and Horowitz argue that autonomous weapons are “brittle” – able to perform well in predictable environments but sometimes failing badly in unanticipated situations. They cite the 2003 Patriot fratricides as an example of the potential consequences if autonomous systems were to fail. They point out that a human was, in fact, “in the loop” in the 2003 Patriot fratricides. While the human operators failed to prevent the fratricides from occurring, they were able to intervene and halt the system’s operation afterward, preventing additional fratricides. This may not be the case with fully autonomous weapons. The consequences of a failure with a fully autonomous weapon could be far more severe.

In such a scenario, a system failure – caused either by a malfunction, unanticipated interaction with the environment, or a cyber attack – raises the frightening prospect of mass fratricide on the battlefield. Similar to the 2010 flash crash, a host of autonomous systems malfunctioning could, in theory, rapidly spiral out of control. In the worst case, the result could be fratricides, civilian casualties, or unintended escalation in a crisis – potentially even a “flash war.”\textsuperscript{56}

In a related assessment of risk and predictability in autonomous weapons, which explored similar case studies as Borrie’s work on normal accidents, Scharre further added the dimension of speed as a complicating factor in potentially undermining stability:

The natural tendency in a competitive environment is toward greater speed, necessitating greater automation, further accelerating the pace of battle. The result could be an unstable situation. Unexpected interactions between autonomous systems or hacking could lead to a “flash war,” where conflicts quickly spiral out of human control.\textsuperscript{57}


\textsuperscript{56} Ibid.


32
Other scholars have deepened the understanding of the applicability of normal accident theory to autonomous weapons. In a 2017 paper, Stephanie Carvin applies an in-depth analysis of normal accident theory, its critics, and the countervailing high-reliability theory to autonomous weapons. She concludes “there is every reason to believe that [lethal autonomous weapon systems] will be prone to normal accidents, despite strict control over their use.”

Matthijs Maas has generalized these same concerns to AI systems more broadly and concludes that normal accidents with AI systems are “an inevitability.” Furthermore, Maas points out that these problems are exacerbated in a competitive context. He concludes by suggesting a number of potentially promising regulatory strategies worthy of further exploration, including:

- ways to change how the system is used by users—to find, wherever possible, ways to reduce either the coupling, or the opacity of the AI; to limit the system’s autonomy (or speed) when deployed outside of intended environments. … heterogeneity in deployed AI architectures and networked systems (to insulate systems from flash crashes) … [and] research into ‘explainable’ AI and AI safety more broadly.

Maas also suggests additional research into ways to mitigate risk factors without hampering performance and “organizational and technological innovation which might better promote graceful failures.” Maas acknowledges that there is significantly more work to be done in identifying technological or organizational solutions to normal accidents with AI systems, but nevertheless clearly outlines the problem and proposes a number of avenues for exploring potential solutions.


61 Maas, “Regulating for ‘normal AI accidents’ – Operational lessons for the responsible governance of AI deployment.”
Jürgen Altmann and Frank Sauer have a similar outlook and have made the most in-depth argument for how autonomous weapons could undermine crisis stability. Their central concern is that the combination of speed and complexity could lead to unanticipated interactions between autonomous systems. In a 2017 article in *Survival*, they paint a picture of how this could occur due to interactions between swarms, although cooperative autonomy (aka swarming) is not necessary for such an interaction to occur.\(^{62}\) Their analysis holds true for any situation that would involve adversarial autonomous systems interacting at high speeds in a very complex environment.

One such swarm-combat situation could be a severe political crisis in which adversaries believe that war could break out. With swarms deployed in close proximity to each other, control software would have to react to signs of an attack within a split-second time frame – by evading or, possibly, counter-attacking in a use-them-or-lose-them situation. Even false indications of an attack – sun glint interpreted as a rocket flame, sudden and unexpected moves of the adversary, or a simple malfunction – could trigger escalation.

The nature of military conflict is such that these kinds of interactions could not be tested or trained for beforehand. In addition, it is, technically speaking, impossible to fathom all possible outcomes in advance. Clearly, the interaction of swarms, if fully autonomous, would be unpredictable, and could potentially result in an escalation from crisis to war, or, within armed conflict, to higher levels of violence.\(^{63}\)

Altmann and Sauer observe that “human reasoning” played a major factor in de-escalating crises and restraining from responding to accidents or false alarms during the Cold War. They raise concerns that autonomous weapons that delegated lethal decision-making authority to machines in crises would discard “tried and tested mechanisms for double-checking and reconsideration that allow humans to function as fail-safes or circuit-breakers.”\(^{64}\) They conclude that:

---


\(^{64}\) Ibid, 129.
This, in combination with unforeseeable algorithm interactions producing unforeseeable military outcomes, increases crisis instability and is unpleasantly reminiscent of Cold War scenarios of accidental war.\footnote{Ibid, 129-130.}

Thus, autonomous weapons potentially present multiple complicating factors that, in combination, could cause small incidents to rapidly spiral into significant accidents. These factors include: a high degree of complexity creating susceptibilities to unanticipated behavior; brittle automation that is not able to flexibly respond to novel events; potentially a lack of communication with human operators; and high speed interactions. In various forms, a number of scholars – Haas, Payne, Velez-Green, Borrie, Scharre, Horowitz, Altmann, and Sauer – have tied this potential for unanticipated accidents to not only have tactical effects, but also potentially operational and strategic effects undermining stability.

Other scholars have considered autonomous weapons in broader concepts of stability, including offense-defense balance and arms race stability. Jean-Marc Rickli has weighed the potential of autonomous weapons to affect the offense-defense balance and threshold for the use of force. Rickli has suggested that autonomous weapons might lower the threshold for the use of force by reducing the risk of loss of life on the side of the attacker.\footnote{Rickli, “The Impact of Autonomous Weapons Systems on International Security and Strategic Stability,” in Quentin Ladetto, ed., Defence Future Technologies: What We See on the Horizon (Thun: Armasuisse, 2017), 62, \url{https://deftech.ch/What-We-See-On-The-Horizon/armasuisseW%2BT_Defence-Future-Technologies-What-We-See-On-The-Horizon-2017_HD.pdf}.} Autonomous weapons also could potentially undermine first-strike stability, he has suggested, through swarming tactics to saturate defenses, which could tilt the offense-defense balance in favor of the offense.\footnote{Rickli, “The Impact of Autonomous Weapons Systems on International Security and Strategic Stability,” 62-63.} Payne similarly argues that AI will benefit the offense for multiple reasons: the increased speed and precision of machines and the ability of swarms to disperse then re-concentrate in attack.\footnote{Payne, “AI, warbot;” Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?”, 26; and Payne, Strategy, Evolution, and War, 179.} The combination of these attributes, he argues, “swings the logic of strategy decisively towards attack.”\footnote{Payne, “AI, warbot.”}
Rickli argues that the advantage conferred to the offense would undermine stability:

In an international environment that favors the offensive, the best strategy to counter the offensive use of force is one that relies on striking first. It follows that strategies of pre-emption are very likely to become the norm if [lethal autonomous weapon systems] are developed. Striking first before being attacked will provide a strategic advantage.70

Payne agrees, stating that military AI “creates a powerful spur for moving first in any crisis.”71

The fact that autonomous weapons provide an advantage would also incentivize countries to attempt to acquire them first, Rickli says, thus provoking an arms race in autonomous weapons. Thus, in Rickli’s estimation, autonomous weapons have the potential to undermine stability in multiple ways. They “favor the offense in the offense-defense balance, and thus have the potential to promote arms races.”72

Rickli also considered the prospect of “artificial super intelligence” that would be far superior to humans, concluding that marrying such a technology with autonomous weapons was a “worrying prospect.” Rickli concluded that based on all of these concerns associated with stability, “the international community should be very careful when considering the development of [lethal autonomous weapon systems].”73

Jürgen Altmann and Frank Sauer have raised similar concerns to Rickli about offense-defense balance and arms race stability. They similarly conclude that uninhabited vehicles would lower the cost of aggression by sparing soldiers’ lives and that swarming would favor the offense.74 Altmann and Sauer highlight “hard-to-defend-
against autonomous swarms” in particular as a destabilizing technology.\textsuperscript{75} Altmann and Sauer also identify that autonomous weapons are at a high risk of an arms race because the technology draws so heavily on dual-use commercial advances in autonomy. This leads them to conclude that if one state were to develop autonomous weapons, “it would be comparably easy – and thus very likely – that others would follow suit.” Because of the high risk of proliferation if one country were to develop autonomous weapons, they argue that “the development of [autonomous weapon systems] could well trigger a destabilizing arms race.”\textsuperscript{76} Due to these concerns, Altmann and Sauer join Rickli and Haas in suggesting that nations should consider cooperative measures to avoid the dangers of autonomous weapons, and they examine several arms control options.

Prospects for arms control

The calls by Haas, Rickli, Altmann, and Sauer for international cooperation to mitigate these risks has ample precedent. While calls for a ban on the basis that autonomous weapons are illegal may be tautologically useless from a purely legal point of view (if a weapon is illegal then it is already banned), there is ample precedent for banning or regulating weapons for strategic stability reasons. Arms control has long been one tool in policymakers’ toolkit to reduce instability. By the late 1950s, U.S. strategists had begun to consider arms control as one possible solution to the problem of strategic stability, in addition to unilaterally reducing U.S. vulnerabilities to a surprise attack. By 1958, the U.S. and U.S.S.R. had begun discussing possible ways to cooperate to reduce the threat of a surprise attack. Over the course of the Cold War, many weapons were regulated or banned not for moral or ethical reasons, but because of their potential to undermine strategic stability. These have included bans on placing nuclear weapons on the seabed (Seabed Treaty), in space (Outer Space Treaty), or weapons of any kind on the moon (Outer Space Treaty); prohibitions on intermediate-range missiles (INF Treaty) and anti-ballistic missile weapons (ABM Treaty); and soft norms against kinetic anti-satellite weapons and neutron bombs.

In fact, there is a specific precedent under U.S. law for concern regarding the strategic stability implications of autonomous weapons. During development of the Strategic Defense Initiative (aka “Star Wars”) in the 1980s, Congress passed a

\textsuperscript{75} Ibid, 131.

resolution in the 1988-89 National Defense Authorization Act prohibiting strategic ballistic missile defenses that “initiate the directing or damaging of lethal fire except by affirmative human decision at an appropriate level of authority.”

In spite of these precedents, some have criticized calls for a ban on autonomous weapons on the ground that it is not feasible. In responding to the call by AI scientist Stuart Russell for a ban, Evan Ackerman has argued that a ban could not succeed because the underlying technology that enables autonomous weapons is so ubiquitous. “The barriers keeping people from developing this kind of system are just too low,” Ackerman argues. “There’s simply too much commercial value in creating quadcopters (and other robots) that have longer endurance, more autonomy, bigger payloads, and everything else that you’d also want in a military system.” It seems reasonable to assume that the accessibility of a technology affects the feasibility of implementing a ban; more widely-available technologies would require greater numbers of potential users to cooperate to effectively implement a ban. Ban feasibility is therefore another important dimension to examine.

Some scholars have examined the feasibility of arms control for autonomous weapons. Rebecca Crootof and Sean Watts both examined historical attempts to ban weapons and identified common characteristics of successful weapons bans. Crootof identifies eight traits often associated with successful weapons bans and concludes that only one of these applies in the case of autonomous weapons. The others are either “inconclusive or currently weigh against the likelihood of a successful ban.” Watts identifies six criteria that lead weapons to being either “regulation tolerant” or “regulation resistant.” Watts concludes that “while autonomous weapon systems presently demonstrate characteristics of both regulation tolerance and resistance, on

77 “No agency of the Federal Government may plan for, fund, or otherwise support the development of command control systems for strategic defense in the boost or post-boost phase against ballistic missile threats that would permit such strategic defenses to initiate the directing of damaging or lethal fire except by affirmative human decision at an appropriate level of authority.” 10 U.S.C. 2431 Sec. 224.


80 Ibid, 1891.

balance, considerations of [regulation] resistance … prevail and indicate a low likelihood that States will in the immediate future consent to a meaningfully-ratified specific regulation or a preemptive ban.” 82

Jürgen Altmann and Frank Sauer come to slightly different conclusions. They view a non-proliferation regime that would attempt to limit access to the underlying technology that would enable autonomous weapons to be not only “futile” but in fact “severely misguided” given the vast societal benefits of artificial intelligence and autonomy. 83 They are less pessimistic, however, on the feasibility of regulating “certain applications of specific technologies for military purposes” and point to the blinding laser ban as an example. 84 Altmann and Sauer acknowledge the challenges in verifying an arms control treaty, “since the software could be changed back within minutes after inspection,” but suggest that ex post solutions to verify compliance after an incident using forensic records might be a solution, perhaps leveraging new technologies such as “manipulation-resistant database solutions such as blockchain.” 85 Ultimately, they conclude that a ban on autonomous weapons “would not be easy to accomplish,” but they point out, “arms control almost never is.” 86 They argue that the “long list of concerns” surrounding autonomous weapons – legal, ethical, and strategic – weigh in favor of nation-states working to “collectively stop the race toward full autonomy in weapon systems.” 87

These analyses by Crootof, Watts, Altmann, and Sauer are illuminating and insightful and collectively suggest that regulating autonomous weapons would be extremely challenging. Surprisingly, however, aside from their work there is a general dearth of theories for why some weapons bans are successful and unsuccessful. There are over 40 historical examples of successful and unsuccessful weapons bans from which to draw conclusions, though. Building on the work of Crootof and Watts, these historical examples can help explain why some bans succeed and others fail and provide insights on the prospects for banning or regulating autonomous weapons.

82 Ibid, 3.
85 Ibid, 135.
86 Ibid, 135.
87 Ibid, 134.
Summary of literature

Autonomous weapons are one application of artificial intelligence in the military sphere, and other uses of AI may also affect stability in significant ways. These include, but are not limited to, AI-enabled robotic systems that improve intelligence collection, automated intelligence analysis, AI-enabled strategic decision-making, automation in command-and-control, and robotic systems that might enhance or undermine nuclear deterrent capabilities. As a general-purpose enabling technology, much like electricity or the internal combustion engine, AI will have a wide range of military applications. There is already a robust and rapidly growing literature exploring broader applications of AI to stability, and nuclear stability in particular. This thesis tackles one specific application of AI, autonomous weapons, and their implications for stability.

Collectively, the handful of articles that have been written on autonomous weapons and stability pose less an established field of literature than an initial set of arguments and hypotheses. Through articles and presentations, a number of scholars have begun to raise questions about the effect of autonomous weapons on number of dimensions of stability: crisis stability, offense-defense balance, and arms race stability. Concerns about crisis stability have been most clearly articulated by Altmann and Sauer, who have argued that their propensity for unanticipated actions combined with a lack of human control could lead to unintended escalation.

These scholars’ concerns take on particular salience in the midst of ongoing international discussions on autonomous weapons, including a growing number of voices calling for a ban on their development and use. A dominant theme across all these articles is the suggestion that discussions on autonomous weapons ought to expand beyond legal and ethical issues to also include a consideration of their impact on things like stability. This thesis contributes to that nascent literature.

Research Questions

This thesis examines the stability dimensions of autonomous weapons and prospects for risk mitigation. The central question is:

_Do autonomous weapons pose challenges to stability and, if so, what are feasible risk mitigation measures that nations could take?_

Specific sub-questions include:

I. Autonomy and autonomous weapons
   - What is the nature of autonomy? What is an autonomous weapon?
   - How has autonomy historically been incorporated into weapons?
   - How do autonomous weapons differ from existing weapons that retain more direct human control over target selection and engagement?

II. Trends in autonomy in weapons
   - What kinds of autonomy are being incorporated into next-generation weapons being built today?
   - What do trends suggest for how autonomy is likely to be used in future weapons?
   - How does autonomy in weapons intersect with cyberspace and cyber weapons?

III. Stability and autonomous weapons
   - What is the concept of stability? What are different kinds of stability? What conditions contribute to or detract from stability?
   - In what ways might autonomous weapons affect stability? How might autonomous weapons affect crisis stability, escalation control, or war termination?
• Could autonomous weapons, either individually or in combination with adversary autonomous weapons, lead to a condition of “fragile stability” where crises rapidly escalate out of control of human commanders?

• Could autonomous weapons increase stability by increasing predictability in crises?

**IV. Controlling autonomous weapons: risk, predictability, and hazard**

• How does risk, predictability, and hazard change as systems become more autonomous?

• What are the likelihood and consequences of failures or accidents with autonomous weapons compared to existing semi-autonomous weapons?

• What do experiences with existing semi-autonomous or supervised autonomous weapons, such as the Patriot air defense system and Aegis combat system, suggest about risk and control in autonomous weapons?

• What do experiences with non-military autonomous systems in competitive or unconstrained environments suggest about control over autonomous weapons?

• How could militaries ensure predictable and fail-safe operations in realistic operating conditions, including under conditions of degraded or denied communications, adversary hacking, and spoofing?

• How do existing theories for safety in complex systems – normal accident theory and high-reliability organizations – apply to autonomous weapons?

• What do historical precedents for dangerous technologies, such as nuclear weapons, suggest about militaries’ ability to effectively manage these risks?

**V. Machine learning systems**

• How do recent advances in machine learning change these risks?

• How would the development of artificial general intelligence change these risks?
VI. Operational factors surrounding autonomous weapons

- How valuable are fully autonomous weapons likely to be, relative to semi-autonomous weapons that would retain a human “in the loop”?
- What might the consequences be of accidents with fully autonomous weapons, in physical space or cyberspace?

VII. Risk mitigation: policy options and historical precedents

- What arms control, transparency, or confidence-building measures could states use to mitigate the potential risks of autonomous weapons, unilaterally or collectively? How effective would these measures be in reducing risk, if successful?
- What lessons can be learned from past historical experiences with weapons that were seen as potentially destabilizing? What parallels exist today with autonomous weapons?
- What do historical examples suggest about the feasibility of various policy options?

Methodology

Examining the stability implications of autonomous weapons is particularly challenging because they are potential future weapons, not weapons that exist today (with some rare exceptions). This makes even understanding the nature of autonomous weapons difficult, not to mention the second and third order effects of their proliferation and conflict dynamics. However, these problems are inherent in the nature of examining emerging technologies. Concern that a new technology might be destabilizing to international security is a reason for preemptive study. One need not wait (and should not wait) until nuclear war breaks out to think about stability dynamics with nuclear weapons, for example. Fortunately, there is a wide body of research in related fields and available case studies to help grapple with these questions.
Over 80 specific technology examples are used from military and non-military settings to better understand autonomous systems’ performance in complex, real-world settings.

- **Stability and escalation dynamics** have been thoroughly explored in the context of other weapons, particularly nuclear weapons. The literature is extensive, stretching decades back through the Cold War. This field of study provides a robust intellectual framework to leverage when considering autonomous weapons. 89

- **Historical uses of automation in weapons.** Automation is used for a wide variety of functions in existing weapons including tracking and identifying targets, prioritizing targets, deciding when to fire, and homing in on targets once launched. Twenty-seven different weapon systems are examined, selected to illustrate different applications of automation:

• **Autonomy in next-generation military systems.** Militaries are pushing the boundaries of autonomy in developmental systems around the globe. Twenty-two systems, chosen to be illustrative of a range of applications, are examined:

  o Autonomous Aerial Cargo/Utility System (AACUS) (U.S.)
  o Robot boat swarms (U.S.)
- Anti-submarine warfare Continuous Trail Unmanned Vessel (ACTUV) (U.S.)
- Armed ground robotic vehicles (non-state groups)
- Aura drone (India)
- Brimstone missile (UK)
- Collaborative Operations in Denied Environments (CODE) program (U.S.)
- B.A.E. ESGRUM unmanned surface vessel (USV) (Ecuador)
- Guardium unmanned ground vehicle (Israel)
- Modified armed drones (non-state groups)
- nEUROn drone (France)
- Platform-M unmanned ground vehicle (Russia)
- Protector USV (Israel)
- RQ-170 drone (U.S.)
- SGR-A1 sentry gun (Republic of Korea)
- Sharp Sword drone (China)
- Skat drone (Russia)
- T-14 Armata tank (Russia)
- Taranis stealth combat drone (UK)
- Target Recognition and Adaptation in Contested Environments (TRACE) program (U.S.)
- Uran-9 unmanned ground vehicle (Russia)
- X-47B prototype drone aircraft (U.S.)

- **Cyber systems.** Automation has long been used in both offensive and defensive cyber systems. Four exemplar systems, and their degree of automation, are examined:
  - Conficker
  - Internet Worm of 1988
  - Mayhem
  - Stuxnet

- **Non-military autonomy and artificial intelligence (AI).** This includes the use of autonomy in self-driving cars, commercial airline autopilots, and games such as
chess or Go. Recent advances in novel methods of AI computation such as deep learning neural networks raise new opportunities and challenges, for example. Specific examples include:

- Automobiles
- Airplane autopilots
- Commercial-grade quadcopters (e.g., DJI Spark)
- Open-source neural network libraries (e.g., Keras)
- IBM Watson Jeopardy Challenge
- AI performance in strategy games (e.g., chess, Go)
- Deep neural networks, including for object identification

- **Normal accidents, predictability, and safety in complex highly automated systems.** The competing theories of normal accident theory and high-reliability organizations provide a grounding for analysis of safety in complex systems. Thirteen examples of accidents or high-reliability are examined to understand the conditions that give rise to either accidents or high-reliability across a range of industries:


• **Automation as a factor in safety and risk in military settings.** In the military setting, automation has been a factor in both accidents and improved safety, similar to other industries. Six examples are examined in order to better understand the application of normal accident theory and high-reliability organizations to the use of automation in military settings:

  o Automatic ground collision avoidance system (Auto-GCAS)
  o Software-based limits of F-35 flight controls
  o 2007 malfunction of a South African anti-aircraft gun
  o 2010 loss of control of a U.S. Navy Fire Scout drone
  o 2011 loss of control and crash of a U.S Air Force RQ-170 drone
  o 2017 loss of control and crash of a U.S. Army Shadow drone

• **Autonomous systems used in competitive environments** impose an additional layer of challenges, and there are important non-military case studies that can be mined for lessons, especially in the financial sector. Five cases are examined:

  - Amazon automated book pricing algorithm price war
  - 2012 Knightmare on Wall Street incident
  - 2010 stock market Flash Crash
  - Post-Flash Crash risk mitigation measures (e.g., circuit breakers)
  - Subsequent mini-flash crashes

• **Operational experience with supervised autonomous weapons.** Militaries around the globe have been using supervised autonomous weapons for air and missile defense for decades, providing a depth of operational experience about one type of autonomous weapons. The U.S. Army experience with the Patriot system and the U.S. Navy’s experience with the Aegis system provide an ideal comparative study, since both systems perform the same mission (air and missile defense) and are built by the same defense contractor but are used differently by the two services. Their culture, operational experience, and safety track record is examined, including:

  - 1988 *USS Vincennes* incident (Iran Air Flight 655 shootdown)
  - 2003 Patriot Tornado shootdown

---


- 2003 Patriot F-18 shootdown
- Patriot operational experience in ballistic missile defense during the 2003 Iraq invasion
- Aegis safety track record post-1988

- **Nuclear weapon safety and risk** is a particularly useful touchstone for thinking about risk and safety with autonomous weapons. Despite the clear danger of nuclear weapons, there have been a frightening number of near-miss incidents throughout the nuclear age. Some of these incidents could have led to the loss of control of a nuclear weapon or its detonation. Others could have conceivably led to a nuclear exchange between superpowers. Nuclear weapon safety has been the subject of much analysis, in particular Scott Sagan’s *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons.* A recent Chatham House report details 13 near miss nuclear use incidents over the past 60 years.

- **Attempts at arms control** – some successful and some unsuccessful – stretch back millennia. A number of ancient Indian texts – the Dharmaśāstras, Mahābhārata, and Laws of Manu – prohibit barbed or poisoned arrows and other weapons. As technologies have developed they have been banned, restricted, or permitted for a variety of reasons. Crossbows were initially banned (for use against Christians) by two papal decrees in the 12th Century. Japan effectively banned firearms for roughly 250 years from the 17th to mid 19th Century, while similar attempts in England in the 16th Century failed. Attempts were made at the turn of the twentieth century to regulate submarines and prohibit aerial attacks on cities. Chemical weapons were banned, as were explosive, expanding, or incendiary projectiles. Today, a host of

---


laws, policies, and normative principles govern a wide range of weapons and their uses.\textsuperscript{99} Treaties exist banning chemical and biological weapons, cluster munitions, land mines, blinding lasers, weapons of mass destruction in space or on the sea floor, weapons of any kind on the moon, and using the environment as a weapon. Nations have explored normative frameworks or codes of conduct for space and cyberspace, and the United States in particular has sought to establish and solidify a norm against kinetic anti-satellite attacks that generate space debris.\textsuperscript{100} In total, over 40 examples of successful and unsuccessful past efforts at regulating weapons provide a rich reservoir of case studies to draw upon to understand the feasibility of regulating autonomous weapons.

- **Interviews from experts** in a variety of fields are used in addition to these other sources. These include experts in human-machine factors engineering, machine learning, human and machine cognition, automation in safety-critical systems, and military operation of automated systems. These interviews in particular bring to light new information about the operation of complex military autonomous weapon systems: the U.S. Navy Aegis combat system and the U.S. Army Patriot air and missile defense system. Interviews with technical and operational experts in the Navy and Army familiar with these systems present new information on the systems’ operation, the cultures surrounding their use, safety procedures, and prior accidents.

Any examination of the stability dimensions of autonomous weapons is prospective, but the problem, by its very nature, requires prospective thinking and analysis. The risks of weapons that are potentially destabilizing to international security should be considered before their deployment. Haas, Borrie, Rickli, Sauer, Altmann, Payne, Velez-Green, and others raise important questions that should be considered and which have not, to date, been examined in sufficient detail. Despite the lack of direct experimental evidence on the effects of fully autonomous weapons, this examination need not proceed in the dark.


As outlined above, there is a robust existing literature and ample case studies in related fields that can be leveraged to make meaningful predictions about the performance of autonomous weapons in real world environments. These include the existing literature on stability and crisis escalation; historical uses of autonomy in weapons; future trends in military robotic systems; autonomy in cyber systems; autonomy and AI in non-military applications; normal accidents and complex systems in non-military settings; military experience with automation as a factor in risk and safety; autonomous non-military systems in competitive environments; prior military operational experience with supervised autonomous weapons; nuclear weapons safety; and arms control. In addition to evaluating the literature in these fields, this project includes targeted interviews with experts in these fields. Leveraging this diverse and deep body of information and case studies, this thesis predicts how autonomous weapons will affect stability and what measures are feasible and appropriate in response.

Original Contributions

This thesis builds on the nascent literature surrounding autonomous weapons and the questions that scholars have raised about their effect on stability and furthers that discussion with new contributions. This thesis:

1) Presents a novel typology of autonomy as a three-dimensional characteristic of a system, rather than a single spectrum of autonomy (Part I). This presentation helps to illuminate aspects of autonomy, both for autonomous weapons and in other applications.

2) Brings to bear over 50 case studies of autonomous military and cyber systems in use or currently in development around the globe (Parts I & II). This is the most detailed series of case studies of autonomy in weapon systems compiled to-date.

3) Applies existing theories of stability to autonomous weapons in order to understand their possible effects in increasing or decreasing stability in a more comprehensive manner than has previously been done elsewhere. Many scholars have raised concerns about the effect of autonomous weapons on stability, but this thesis classifies these concerns, as well as new ones, in a comprehensive manner (Part III). In total, nine mechanisms by which
autonomous weapons could affect stability are identified, four of which increase stability and five undermine stability.

4) Using the theoretical framework of normal accident theory and high-reliability organizations, analyzes over 25 case studies of military and non-military uses complex or autonomous systems to better understand risk, predictability, and control in real-world environments (Part IV). Additionally, this analysis applies high-reliability and normal accident theories to complex systems in adversarial environments and presents a set of criteria for when high-reliability theory is likely to prevail and when normal accident theory is likely to prevail, a novel application of these theories. These criteria make predictions about the likely safety record of complex autonomous systems in adversarial environments and is testable through future observations.

5) Presents original in-depth case studies of two military weapon systems, the U.S. Army Patriot air and missile defense system and the U.S. Navy Aegis combat system, including interviews with experts familiar with both systems (Part IV). This includes new information on the operation of both of these weapon systems, the cultures surrounding their use, and safety procedures that have not been presented elsewhere, drawing on interviews with U.S. defense professionals who work on these weapon systems. These include the U.S. Army engineering psychologist who led the Army’s post-mortem assessment of the 2003 Patriot fratricides and the commander of the Navy’s Aegis Training and Readiness Center, which trains all Aegis-qualified officers and enlisted sailors. The findings of these case studies have direct bearing on the likelihood of building safe and controllable future autonomous weapon systems.

6) Analyzes the operational value and risks of autonomous weapons, an under-explored area to-date (Part VI).

7) Presents a comprehensive dataset of over 40 historical attempts to ban or otherwise restrict weapons technologies, dating to ancient India (Appendix D). This is the most comprehensive dataset of historical attempts to ban weapons that has been compiled to-date.
8) Presents a novel framework for the likelihood of attempted weapons bans’ success (Part VII). While other scholars have previously presented criteria that influence the likelihood of a ban’s success, this thesis builds on this existing work to present a new and different framework. This framework is applicable to any weapon, not just autonomous weapons. Additionally, this framework has predictive value and is testable by observing the success or failure of future weapons bans or existing bans under future conditions.

9) Explores four possible regulatory approaches to autonomous weapons and makes testable predictions about their likelihood of success based on historical attempts to ban weapons (Part VII).

This thesis contributes to the field by: applying existing theories to new case studies; presenting new case studies with new information; compiling and classifying data in new ways (effect of autonomous weapons on stability, historical weapons bans); and presenting new frameworks for understanding various issues (nature of autonomy, operational value and risk of autonomous weapons, likelihood of weapons bans’ success). These contributions help to understand the impact of autonomous weapons on stability. They also contribute more broadly to the study of automation, risk, and hazard in adversarial settings and to the study of attempts to limit or restrict weapons. This thesis presents testable hypotheses on the likelihood of accidents with complex autonomous systems in adversarial environments and on the likely success of weapons bans under certain conditions.

---

Outline of the Report and Preview of Conclusions

This report will proceed in the following manner:

Part I examines the nature of autonomy and autonomous weapons through 27 case studies of how autonomy has historically been used in various military and non-military systems.

Part II explores over 25 case studies of next-generation weapon systems and cyber systems with increasing degrees of autonomy that are in development by countries around the globe. Nations’ transparency regarding the details of new weapons programs varies considerably, which influences which cases are available for study. Part II concludes that the overall trend is towards increasing autonomy and, while no nation has clearly demonstrated an intent to develop autonomous weapons, the technology will become increasingly accessible.

Part III examines the intersection of autonomous weapons and stability and develops nine mechanisms by which autonomous weapons could increase or decrease stability. Part III concludes that the most significant stability issues for autonomous weapons are their effect on crisis stability caused by unintended behaviors. The net effect of these features could be to undermine crisis stability, escalation management, and war termination.

Part IV evaluates concerns about predictability, controllability, and hazard of autonomous weapons via over 25 case studies of military and non-military systems. These include case studies uses of autonomy and artificial intelligence in adversarial settings. Part IV concludes that predictability and controllability are a significant problem in complex, highly automated systems for a variety of reasons and that these risks are exacerbated in adversarial settings. Additionally, Part IV identifies that while there are some cases in which organizations are able to operate complex, hazardous systems with few accidents, these conditions do not apply to the case of fully autonomous weapons.

Part V evaluates these risks in light of recent advances in artificial intelligence and machine learning and hypothetical future advances towards human-like artificial general intelligence. Part V concludes that accident risks are exacerbated, rather than mitigated,
by current forms of machine learning due to their counterintuitive inhuman properties and vulnerabilities to spoofing attacks, bias, and other safety concerns. Achieving human-like artificial general intelligence, if possible someday, would mitigate some risks but likely raise a host of far more grave safety concerns.

Part VI explores the military operational value and dangers of autonomous weapons, including pathways for escalation. Part VI concludes that while semi-autonomous and supervised autonomous weapons may be appropriate for most military settings, there are some situations in which fully autonomous weapons would be valuable. Part VI also concludes that the risks of unintended escalation are real, and that the potential for high-speed unintended interactions (“flash war”) are most severe in cyberspace.

Part VII examines prospects for risk mitigation and arms control, including historical experiences with over 40 prior successful and unsuccessful attempts at regulating other weapons. Part VII concludes that for a number of reasons a comprehensive ban on autonomous weapons is unlikely and outlines three other alternative regulatory regimes that have a higher likelihood of success.
PART I: AUTONOMY AND AUTONOMOUS WEAPONS

This section examines the nature of autonomy and autonomous weapons by exploring how autonomy has historically been used in various military and non-military systems. It will first explain why autonomy is important in military systems, tracing the history of the recent explosion of military robots and drones. It will then examine the nature of autonomy by drawing on examples on autonomous systems in other industries, such as thermostats, self-driving cars, and robot vacuum cleaners. Finally, this section will trace the history of autonomy in weapons to-date and how autonomous weapons differ from existing weapons that retain more direct human control over target selection and engagement.

The AI and Robotics Revolution

Just as robots are transforming a range of industries, from self-driving cars to robot vacuum cleaners and caretakers for the elderly, they are also transforming war. Over 90 countries have aerial drones today, along with many non-state groups. Scores of countries are expanding their arsenals of air, ground, and maritime robots, and with each generation they are becoming more autonomous.

Robots have many battlefield advantages over traditional human-inhabited vehicles. Unshackled from the physiological limits of humans, uninhabited (“unmanned”) vehicles can be made smaller, lighter, faster, and more maneuverable. They can stay out on the battlefield beyond the limits of human endurance, for weeks, months, or even years at a time. They can take more risk, opening up tactical opportunities for dangerous or even sacrificial missions without risking human lives.

Robots have one major disadvantage, however. By removing the human from the vehicle, they lose the most advanced cognitive processor on the planet: the human brain. Most military robots today are remotely controlled, or tele-operated, by humans. They depend on fragile communication links that can be jammed or disrupted by environmental conditions. Without communications, robots can perform simple tasks, but their capacity for autonomous operation is limited. Greater intelligence and autonomy is desirable, in part, to enable more capable robotic operations in communications denied environments.
The robotics revolution

The U.S. military deployed thousands of air and ground robots to meet urgent needs in Iraq and Afghanistan. Spending on uninhabited aircraft, which had hovered around the $300 million per year mark in the 1990s, skyrocketed after 9/11, increasing six-fold to over $2 billion per year by 2005.\textsuperscript{102} By 2011, annual spending on drones had swelled to over $6 billion per year, over twenty times pre-9/11 levels.\textsuperscript{103} The Department of Defense (DoD) had over 7,000 drones in its fleet. The vast majority of them were smaller hand-launched models, but also included large aircraft like the MQ-9 Reaper and RQ-4 Global Hawk.\textsuperscript{104}

While drones were not a new technology – they had been used in a limited fashion in Vietnam – the overwhelming crush of demand for them from warfighters was new.\textsuperscript{105} Drones give commanders a low-cost, low-risk way to put eyes in the sky. While in later years drones would become associated with “drone strikes,” it is their capacity for persistent surveillance, not dropping bombs, that makes them unique and valuable to the military.

Robots are equally if not more important on the ground. Driven in large part by the rise of improvised explosive devices (IEDs), DoD deployed over 6,000 ground robots to Iraq and Afghanistan.\textsuperscript{106} Small ground robots like the iRobot Packbot allowed troops to disable or destroy IEDs without putting themselves at risk.

Increased autonomy

In 2005, DoD began publishing a series of “roadmaps” for future unmanned system investment. The 2005 roadmap was focused on aircraft, but subsequent roadmaps in 2007, 2009, 2011, and 2013 included air and maritime vehicles as well.

\textsuperscript{102} Office of the Secretary of Defense, “Unmanned Aircraft Systems Roadmap, 2005–2030,” August 4, 2005, 37, Figure 2.6-1, \url{http://fas.org/irp/program/collect/uav_roadmap2005.pdf}.


Each roadmap looked forward 25 years, outlining technology goals to help inform future investments by government and industry. They covered sensors, communications, power, weapons, propulsion, and other key enabling technologies. A dominant theme across all the roadmaps was autonomy. The 2011 roadmap perhaps summarized it best:

For unmanned systems to fully realize their potential, they must be able to achieve a highly autonomous state of behavior and be able to interact with their surroundings. This advancement will require an ability to understand and adapt to their environment, and an ability to collaborate with other autonomous systems.\(^{107}\)

Autonomy is the cognitive engine that power robots. Without autonomy, robots are only empty vessels that depend on communication links to remote human controllers to perform tasks.

Today’s military robots are simple and largely remote controlled. In Iraq and Afghanistan, the U.S. military has operated in a relatively “permissive” electromagnetic environment where adversaries generally did not have the ability to jam communications. This will not always be the case in future conflicts. Major nation-state militaries will almost certainly have the ability to disrupt or deny communications networks, and the electromagnetic spectrum will be highly contested. The U.S. military has ways of communicating that are more resistant to jamming, but these methods are limited in range and bandwidth. The type of drone operations the United States has conducted when going after terrorists – streaming high-definition, full-motion video back to stateside bases via satellites – will not be possible against a major military power. In addition, some environments inherently make communications challenging, such as undersea, where radio wave propagation is hindered by the water. In these situations, autonomy is a must if robotic systems are to be valuable. As AI advances, more sophisticated autonomous robots will be able to conduct more complex missions in more challenging environments independent from human control.

Greater autonomy is also desirable even if there are perfect communications links because of the personnel burden to remotely control individual robots. Thousands of robots require thousands of people to control them if each robot is remotely operated.

Predator and Reaper drone operations require 7 to 10 pilots to staff one drone “orbit” of 24/7 continuous around-the-clock coverage over an area. Another 20 people per orbit are required to operate the sensors on the drone, and scores of intelligence analysts are needed to sift through the sensor data. In fact, because of these substantial personnel requirements, the U.S. Air Force has a strong resistance to calling these aircraft “unmanned.” The aircraft itself is uninhabited, but there are still humans controlling it and supporting it.

Because the pilot remains on the ground, uninhabited aircraft unshackle surveillance operations from the physical limits of human endurance. Drones can stay aloft for 12 to 28 hours at a time, far longer than a human pilot could remain effective sitting in the cockpit. But remote operation doesn’t change the cognitive requirements on human operators. Humans still have to perform the same tasks, they just aren’t physically onboard the vehicle. The U.S. Air Force prefers the term “remotely piloted aircraft” because that’s precisely what most current drones are. Pilots still fly the aircraft via stick and rudder input, just remotely from the ground, sometimes even half a world away.

Remote control is a cumbersome way to operate. Even under the best of circumstances, there may be communications delays and limited bandwidth. Remote operations limit a military’s ability to effectively capitalize on robotics. Building tens of thousands of cheap robots is not a cost-effective strategy if they require even larger numbers of highly trained (and expensive) people to operate them. Autonomy is a solution. The 2011 DoD roadmap stated:

> Autonomy reduces the human workload required to operate systems, enables the optimization of the human role in the system, and allows human decision making to focus on points where it is most needed. These benefits can further result in manpower efficiencies and cost savings as well as greater speed in decision-making.

Greater autonomy is a consistent theme across DoD documents on robotics, with many of the roadmaps pointing towards a long-term goal of achieving “full autonomy.” The 2005 roadmap looked towards “fully autonomous swarms.” The 2011 roadmap

---

108 Ibid, 45.

articulated four levels of autonomy from human operated (level 1) to human delegated (level 2), human supervised (level 3), and eventually fully autonomous (level 4). The benefit of greater autonomy was the “single greatest theme” in a 2010 report from the Air Force Office of the Chief Scientist on future technology.

Incrementally, engineers are adding to the set of tasks that uninhabited aircraft can perform autonomously on their own, moving step by step towards increasingly autonomous drones. In 2013, the U.S. Navy autonomously landed its X-47B prototype drone on a carrier deck at sea. The only input from a human operator was to direct the aircraft to land. The actual flight onto the carrier deck was done by software. In 2014, the Navy’s Autonomous Aerial Cargo/Utility System (AACUS) demonstrated the ability for an uninhabited helicopter to autonomously scout out an improvised landing area in the field and land on its own. Again, the only input required from a human was to push a button directing the helicopter to land. In 2015, the X-47B drone again made history by conducting the first autonomous aerial refueling, taking gas from another aircraft while in flight.

These are key milestones in building more fully combat-capable uninhabited aircraft. Just as autonomous cars will allow a vehicle to drive from point A to point B without direct human supervision, the ability to takeoff, land, navigate, and refuel autonomously will allow robots to perform tasks under human direction and supervision, but without humans controlling each vehicle movement. This breaks the paradigm of humans directly controlling the robot’s movements, shifting humans into a supervisory role. Human control will still exist, but at a higher level. Humans will command the robot what action to take, and the robot executes all on its own.

The next step is cooperative autonomy, or swarming. Swarming entails multiple autonomous robots that coordinate among themselves to take an action as directed by a human controller. Swarming has many benefits. Swarming can allow militaries to field large numbers of assets on the battlefield with a small number of human controllers. Cooperative behavior can also allow quicker reaction times, so that the swarm can respond to changing events faster than would be possible with one person controlling each aircraft. Swarming robots have potentially significant advantages over adversaries.

---


in mass, coordination, and speed. Because swarms further change the human-machine relationship, with the human commanding the entire swarm which then executes tasks on its own, this raises novel issues of command-and-control.112

In August 2014, the U.S. Navy demonstrated a swarm of small boats on the James River in Virginia.113 The boats were used to simulate a mock strait transit where they protected a high-value Navy ship against potential threats, escorting it through a simulated high-danger area. When directed by a human controller to investigate a potential threat, a detachment of boats moved to intercept and encircle the suspicious vessel. The boats coordinated their actions by sharing information and all their movements were done autonomously. The human controller’s role was simply to direct them to intercept the approaching ship. The five boats worked cooperatively together, although the concept could be scaled up to larger numbers, similar to the aerial drone swarms.114

Robotic swarm boats are highly valuable to the Navy as a potential way to guard against threats to its ships. Robotic boats could intercept suspicious boats further away, putting eyes (and potentially weapons) on hostile boats without putting any sailors at risk.

This kind of application raises important questions about how much autonomy is appropriate in armed robotic systems. In a video released by the Office of Naval Research (ONR) about this demonstration, a .50 caliber machine gun is seen on one of the boats, which the video’s narrator explains could be used to “damage or destroy hostile vessels.” When pressed by reporter following the demonstration, a spokesman for ONR explained “there is always a human in the loop when it comes to the actual engagement of an enemy.” However the spokesman acknowledged, “Under this swarming demonstration with multiple [unmanned surface vehicles], ONR did not study the specifics of how the human-in-the-loop works for rules of engagement.”115

---


114 Ibid.

essence, while the Navy acknowledged that the rules of engagement for the robotic boat would be important, they did not have answer to what those rules of engagement should be.

The Navy’s answer reflects a tension in the military’s pursuit of more advanced robotics. Even as researchers and engineers move to incorporate greater autonomy, there is an understanding that there are – or should be – limits on autonomy when it comes to weapons use. What exactly those limits are, however, is often unclear.

*Pushing humans “out of the loop”*


**Observe, Orient, Decide, Act (OODA) loop**

![OODA Loop Diagram](image)

In the OODA loop paradigm of combat, victory on the battlefield goes to whichever side can complete the observe-orient-decide-act cycle faster.

Pilots talk about an observe, orient, decide, act (OODA) loop in air combat, a cognitive process pilots go through to see and engage enemy aircraft. In the OODA loop paradigm of air combat, victory goes to the pilot who can complete this process faster than the enemy. Understanding the environment, deciding, and acting faster than the enemy allows a pilot to “get inside” the enemy’s OODA loop. While the enemy is still trying to understand what’s happening and decide on a course of action, the pilot has
already changed the situation, resetting the enemy to square one and forcing him or her to come to grips with a new situation.\textsuperscript{116}

If victory comes from completing this cognitive process faster, then automation has significant advantages. The Air Force Flight Plan saw potential for computers to exceed human decision-making speeds:

Advances in computing speeds and capacity will change how technology affects the OODA loop. ... In 2047 technology will be able to reduce the time to complete the OODA loop to micro or nanoseconds. Much like a chess master can outperform proficient chess players, [unmanned aircraft systems] will be able to react at these speeds and therefore this loop moves toward becoming a “perceive and act” vector. Increasingly humans will no longer be “in the loop” but rather “on the loop” – monitoring the execution of certain decisions. Simultaneously, advances in AI will enable systems to make combat decisions and act within legal and policy constraints without necessarily requiring human input.\textsuperscript{117}

This is the logical culmination of an arms race in speed: autonomous weapons that complete engagements all on their own. The Air Force acknowledged the gravity of what it was suggesting might be possible. The very next paragraph continued:

Authorizing a machine to make lethal combat decisions is contingent upon political and military leaders resolving legal and ethical questions. These include the appropriateness of machines having this ability, under what circumstances it should be employed, where responsibility for mistakes lies and what limitations should be placed upon the autonomy of such systems. … Ethical discussions and policy decisions must take place in the near term in order to guide the development of future [unmanned aircraft systems] capabilities, rather than allowing the development to take its own path apart from this critical guidance.\textsuperscript{118}


\textsuperscript{118} Ibid, 41.
The Air Force wasn’t recommending autonomous weapons or even suggesting they were necessarily a good idea. What it was suggesting was that autonomous systems might have advantages over humans in speed, and that AI might advance to the point where machines could carry out lethal targeting and engagement decisions all on their own. If true, then legal, ethical, and policy discussions should take place now in order to shape the development of this technology, the Air Force argued.

At that time, there was no U.S. policy on autonomy. There wasn’t even a formal effort underway. The 2011 roadmap took a stab at an answer, even if it was a temporary one:

Policy guidelines will especially be necessary for autonomous systems that involve the application of force. … For the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems.119

It didn’t say much, but was the first official DoD policy statement on lethal autonomy. Lethal force would remain under human control for the “foreseeable future.” With technology rapidly developing, this answer would not remain sufficient for long. In November 2012, the DoD released a more comprehensive policy, DoD Directive 3000.09, Autonomy in Weapon Systems.120 This policy, which remains in effect today, lays out more detailed guidance for weapons developers. It gives approval for applications that are consistent with how the United States has historically used autonomy in weapons. The policy delineates what these uses of autonomy are, so that there is no confusion among developers. For new uses of autonomy that would be inconsistent with historical patterns, the policy establishes a path for weapons developers to seek approval for development and use. The policy includes the criteria for approval and establishes which DoD officials have approval authority. Nevertheless, the policy does not answer the crucial question as to whether the United States would actually build autonomous weapons. It simply lays out a process for defense officials to grapple with that question. As technology advances, defense officials in the United


120 Disclosure: I was involved in drafting this policy when I worked at the Department of Defense.
States and around the world will inevitably be confronted with the question of whether to delegate lethal engagement authority to machines.

**The Nature of Autonomy**

Autonomy is the freedom for a system (human or machine) to perform a task or function on its own. For autonomous machines, there are three key “dimensions” of freedom:

1) The type of task the machine is performing;
2) The relationship of the human to the machine when performing that task; and
3) The sophistication of the machine’s decision-making when performing the task.

These dimensions are independent, and a machine can be “more autonomous” by increasing the amount of autonomy along any of these dimensions.

1) **Task being performed**

The most important dimension of autonomy is the task being performed by the machine. Not all tasks are equal in their significance, complexity, and risk. A thermostat is an autonomous system in charge of regulating temperature. The fictional military autonomous system Skynet in the Terminator films was given control over nuclear weapons. The complexity of those decisions and the consequences if the machine fails to perform the task adequately are very different. Often, a single machine will perform some tasks autonomously, while humans are in control of other tasks, blending human and machine control within the system. Modern automobiles have a range of automatic or autonomous features: automatic braking and collision avoidance, anti-lock brakes, automatic seat belt retractors, adaptive cruise control, automatic lane keeping, and self-parking. Some automatic or automated functions can be turned on or off by the human user, such as autopilots in commercial airliners. Other automatic functions are always ready and decide for themselves when to activate, such as airbags. Some autonomous systems may be designed to override the human user in certain situations. U.S. fighter aircraft have been modified with an automatic ground collision avoidance system (Auto-GCAS). If the pilot becomes disoriented and is about to crash, Auto-GCAS will
take control of the aircraft and pull up, avoiding the ground. The system has already
saved at least one aircraft in combat, rescuing a U.S. F-16 in Syria.¹²¹

As automobiles and aircraft demonstrate, it is meaningless to refer to a system
as “autonomous” without referring to the specific task that is being automated. F-16s
are still flown by human pilots, but in some instances automation may kick in. Cars are
still driven by humans (for now), but a host of automatic functions can assist the human
in driving or even take control for short periods of time. The system becomes “more
autonomous” as it takes on more tasks, but some degree of human involvement and
direction still exists. “Fully autonomous” self-driving cars can navigate and drive on
their own, but a human is still choosing the destination.

2) Human-machine relationship

For any given task, there are degrees of autonomy even for that specific task. A
system can perform a task in a semi-autonomous, supervised autonomous, or fully
autonomous manner. This is the second dimension of autonomy: the human-machine
relationship.

In semi-autonomous systems, the machine performs a task and then waits for a
human user to take an action before continuing. A human is “in the loop.” Autonomous
systems go through a sense – decide – act loop similar to the military OODA loop, but

¹²¹ Guy Norris, “Ground Collision Avoidance System ‘Saves’ First F-16 In Syria,” Aerospace
avoidance-system-saves-first-f-16-syria.
for semi-autonomous systems the human breaks the loop. A semi-autonomous system can sense its environment and recommend a course of action but cannot carry out the action without human approval.

In supervised autonomous systems, the human is “on” the loop. Once put into operation, the machine can sense its environment, decide on a course of action, and act all on its own. The human user can observe the autonomous system’s behavior and intervene to stop it, if desired. Without any intervention, however, the machine will perform the task on its own.

Fully autonomous systems can sense, decide, and act on their own and the human is not able to intervene, or at least not in a timely manner. The human activates
the machine, but once activated it conducts the task without communication back to the human user. The human is “out of the loop.”

Many machines can operate in different modes at different times. A Roomba that is vacuuming while you are home is operating in a supervised autonomous mode. If the Roomba becomes stuck, you can intervene. If you’re out of the house, then the Roomba is operating in a fully autonomous capacity. If something goes wrong, it’s on its own until you come home.

3) Sophistication/intelligence of machine decision-making

Intelligence is the third dimension of autonomy. More sophisticated, or more intelligent, machines can be used to take on more complex tasks in more challenging environments. People often use terms like “automatic,” “automated,” or “autonomous” to refer to a spectrum of sophistication in machines.

Automatic systems are simple and don’t exhibit much in the way of “decision-making.” They sense the environment and act. The relationship between sensing and action is immediate and linear. It is also highly predictable to the human user. An old mechanical thermostat is an example of an automatic system. The user sets the desired temperature and when the temperature gets too high or too low, the thermostat activates the heat or air conditioning.

Automated systems are more complex. They may consider a range of inputs and weigh several variables before taking an action. Nevertheless, the internal cognitive processes of the machine are generally linear and traceable to the human user, at least in principle. A modern digital programmable thermostat is an example of an automated system. Whether the heat or air conditioning turns on is a function of the house temperature as well as what day and time it is. Given knowledge of the inputs to the system and its programmed parameters, the system’s behavior should be predictable to a trained user.

As systems increase in complexity, people often use the word “autonomous” to refer to systems that are sophisticated enough that their internal cognitive processes are less understandable to the user. The user understands the task the system is supposed to perform, but not necessarily how the system will perform that task. Researchers often refer to autonomous systems as being “goal-oriented.” The human user specifies the goal, but autonomous systems have some flexibility in how they achieve their goal.
An example of a goal-oriented autonomous system is a self-driving car. The user specifies the destination and other goals, such as avoiding accidents. But the human user can’t possibly specify in advance every action the autonomous car must perform. The human doesn’t know where there will be traffic or obstacles in the road, when lights will change, or what other cars or pedestrians will do. The car is therefore programmed with the flexibility to decide when to stop, when to go, and when to change lanes in order to accomplish its goal — getting to its destination safely.

Autonomy doesn’t mean the system is exhibiting free will or disobeying its programming. With every action it takes, the car is still following its programming. The difference is that unlike an automatic system, where there is a simple, linear connection from sensing to action, autonomous systems take into account a range of variables to consider the best action in any given situation. Goal-oriented behavior is necessary for autonomous systems operating in uncontrolled environments that may have obstacles that cannot be precisely predicted in advance. If a self-driving car were operating on a closed track where there weren’t pedestrians or other vehicles, then less sophisticated autonomy might be acceptable. Each movement could be programmed into the car in advance — when to stop, when to go, and when to turn. This car would not be very useful, however, as it could only drive in a simple environment where every necessary action could be predicted. In more complex environments or when performing more
complex tasks, the machine needs the flexibility to make decisions based on the specific situation.

In practice, the line between automatic, automated, and autonomous systems is blurry. At what point does a system cross from being automated to autonomous? Often, the term “autonomous” is used to refer to future systems that have not yet been built. Once they do exist, people often call them “automated.” This is similar to a trend in artificial intelligence where AI is often perceived to be the things that machines cannot yet do. Once a machine conquers a task, then it is merely “software.” Decades ago, many would have considered a computer program that could beat a human at chess to be artificial intelligence. Now that it has been done, chess-playing AI systems are merely computer programs.

The double-edged sword of greater complexity

Greater complexity in autonomous systems is a double-edged sword. On the one hand, more sophisticated systems can undertake more cognitively challenging tasks in more complex environments. The downside to more sophisticated systems, however, is precisely the fact that the human user may not be able to predict its specific actions in advance. The increased autonomy that is a feature can sometimes be a flaw if the human user is surprised in an unpleasant way by the system’s behavior. For simple automatic or automated systems, this is less likely. But as the complexity of the system increases and its internal cognitive processes become less understandable to the human user, then it may be harder for the human to correctly anticipate what the system may do in a novel situation. This could lead human operators to employ autonomous systems in settings for which they are not appropriate, leading to failures and possibly accidents.

A History of Automation and Autonomy in Weapons

The process of incorporating automation into weapons began in the middle of the nineteenth century as inventors began to apply innovations borne out of the industrial revolution to warfare. The Gatling gun, invented at the start of the American civil war in 1861, used automation to speed up the process of firing. A forerunner of the modern machine gun, the Gatling gun employed automation for loading and firing, allowing more bullets to be fired in a shorter amount of time. In the Gatling gun, the

---

process of loading bullets, firing, and ejecting cartridges was all automatic, provided a human kept turning the crank. The Gatling gun was a significant improvement over Civil War era rifled muskets, which had to be loaded by hand through the muzzle in a lengthy process. Well-trained troops could fire three rounds per minute with a rifled musket.\textsuperscript{123} The Gatling gun fired over three hundred rounds per minute. Four soldiers were needed to operate the Gatling gun, but through automation they could deliver equal firepower as a hundred men.\textsuperscript{124} The result was a tremendous expansion in the destructive power unleashed on the battlefield. The Gatling gun was not an autonomous weapon, but began a long evolution of automating various functions of weapons.

\textit{Automatic weapons: Machine guns}

The next tick in the gears of progress came in 1883 with the invention of the Maxim gun. Unlike the Gatling gun, which required a human to hand-crank the gun to power it, the Maxim gun harnessed the physical energy from the recoil of the gun’s firing in order to power the process of reloading the next round. Hand-cranking was no longer required. Once firing was initiated, the gun could continue firing on its own. The machine gun was born.

Unlike semi-automatic weapons, which require the user to pull the trigger for each bullet, today’s automatic weapons will continue firing so long as the trigger remains held down. Machine guns are “dumb” weapons, however, and still have to be aimed by the user. Once initiated, they can continue firing on their own, but the guns have no ability to sense targets. In the Twentieth Century, weapons designers would take the next step to add rudimentary sensing technologies into weapons – the initial stages of intelligence.

\textit{The first intelligent weapons}

From the first time a human threw a rock in anger until the 20\textsuperscript{th} Century, warfare was fought with unguided weapons. Projectiles – whether shot from a sling, a bow, or a cannon – follow the laws of gravity once released. Projectiles are often inaccurate, and their inaccuracy increases with range. Before guided weapons, destroying the enemy hinged on getting close enough to deliver overwhelming barrages of fire to blanket an area. In World War II, as rockets, missiles, and bombs increased the

\begin{thebibliography}{99}
\bibitem{124} George Knapp, “Rifled Musket, Springfield, Model 1861.”
\end{thebibliography}
range at which combatants could target one another, but not the accuracy, militaries sought to develop precision-guided weapons that could accurately strike targets from long distances.

The first successful precision-guided weapon was the German G7e/T4 Falke (Falcon) torpedo, introduced in 1943. The Falke torpedo incorporated a new innovation: an acoustic homing seeker. Unlike traditional torpedoes that traveled in a straight line and could miss a passing ship, the Falke incorporated a homing seeker to account for aiming errors. After traveling 400 meters from the German U-boat that launched it, the Falke would activate its passive acoustic sensors, listening for any nearby merchant ships. It would then steer toward any ships, detonating once it reached them.\(^{125}\)

The Falke was used by only three U-boats in combat before being replaced by the upgraded G7es/T5 Zaunkönig (Wren), which had a faster motor and therefore could target faster moving Allied navy ships in addition to merchant vessels. Using a torpedo that could home in on targets rather than travel in a straight line had military advantages, but also created complications. Two U-boats were sunk in December 1943 (U-972) and January 1944 (U-377) when their torpedoes circled back on them, homing in on the sound of their own propeller noise. The 400-meter safety limit before the homing mechanism activated was intended to address this problem, but to more fully mitigate against the risks of a homing torpedo, German U-boats also began incorporating a tactic of diving immediately after launch and then going completely silent.\(^{126}\)

The Allies quickly developed a countermeasure to the Wren torpedo. The Foxer, an acoustic decoy towed behind Allied ships, was intended to lure away the Wren so that it detonated harmlessly against the decoy, not the ship itself. The Foxer introduced other problems; it loudly broadcast the Allied convoy’s position to other nearby U-boats, and it wasn’t long before the Germans introduced the Wren II with an improved acoustic seeker.\(^{127}\) Thus launched the arms race in ever more intelligent weapons and improved countermeasures against them.


\(^{126}\) Ibid.

\(^{127}\) “The Torpedoes.” “Torpedoes of Germany.” “TV (G7es) Acoustic Homing Torpedo.”
Semi-autonomous weapons: Precision-guided weapons

The latter half of the 20th century saw the expansion of precision-guided weapons into sea surface, air, and ground combat. Today, they are widely used by militaries around the world in a variety of forms. Precision-guided weapons use automation to correct for aiming errors and help guide the munition (missile, bomb, or torpedo) onto the intended target. Depending on their guidance mechanism, precision-guided weapons can have varying degrees of autonomy.

Some precision-guided weapons have very little autonomy, with the human controlling the aimpoint of the weapon throughout its flight. Command-guided weapons are directly controlled by a human remotely via a wire or radio link. For other weapons, a human operator “paints” the target with a laser or radar and the munition homes in on the laser or radar reflection. In those cases, the human doesn’t directly control the movements of the munition, but does control the weapon’s aimpoint in real time. This allows the human controller to re-direct the munition in flight or potentially abort the attack.

Other precision-guided weapons are “autonomous” in the sense that they cannot be recalled once launched, but the munition’s flight path and target are pre-determined. These munitions can use a variety of guidance mechanisms. Nuclear-tipped ballistic missiles use inertial navigation consisting of gyroscopes and accelerometers to guide the missile to its pre-selected endpoint. Submarine-launched nuclear ballistic missiles use star-tracking celestial navigation systems to orient the missile, since the undersea launching point varies. Many cruise missiles look down to the Earth rather than up to the stars for navigation, using radar or digital scene mapping to follow the contours of the Earth to their pre-selected target. GPS-guided weapons rely on signals from the constellation of U.S. global positioning system satellites to determine their position and guidance to their target. While many of these munitions cannot be recalled or redirected after launch, the munitions do not have any freedom to select their own targets or even their own navigational route. In terms of the task they are performing, they have very little autonomy, even if they are beyond human control once launched. Their movements are entirely predetermined. The guidance systems, whether internal such as inertial navigation or external such as GPS, are designed to ensure the munition stays on path to its preprogrammed target. The limitation of these guidance systems, however, is that they are only useful against fixed targets.

Homing weapons, by contrast, are used to track onto moving targets. By necessity since the target is moving, homing munitions have the ability to sense the
target and adapt to its movements. Some homing munitions use passive sensors to
detect their targets, as the Wren did. Passive sensors listen or observe the environment
and wait for the target to indicate its position by making noise or emitting in the
electromagnetic spectrum. Other homing seekers use active measures, such as radars, to
sense a target. An early U.S. active homing munition was the Bat anti-ship glide bomb,
which had an active radar seeker to target enemy ships.\footnote{Jim Sweeney, “Restoration: The Bat,” \textit{Air & Space Magazine} (January 2002), \url{http://www.airspacemag.com/military-aviation/restoration-the-bat-2925632/}.}

Some homing munitions “lock” onto a target before launch; the munition’s
seeker senses the target before it is launched. Other munitions “lock on” after launch –
they are launched with the seeker turned off; then the seeker activates after it is
launched to begin looking for the target attempting to maneuver away, much like an
attack dog running down its prey.

An attack dog is a good metaphor for a fire and forget homing munition. U.S.
pilots refer to the tactic of launching the AIM-120 AMRAAM air-to-air missile in “lock
on after launch” mode as going “maddog.”\footnote{Air Land Sea Application Center, “Brevity: Multi-Service Brevity Codes,” February, 2002, \url{http://www.dtic.mil/dtic/tr/fulltext/u2/a404426.pdf}, I–19. The brevity code for when the seeker goes active is “Pitbull.”} After the weapon is released, it turns on
its active radar seeker and looks for targets. Like a mad dog in a meat locker, it will go
after the first target it sees. Similar to the problem German U-boats faced with the
Wren, pilots need to take care to ensure that the missile doesn’t track onto friendly
targets. Militaries around the world use tactics, techniques, and procedures (“TTPs” in
military parlance) to avoid homing munitions turning back on themselves or other
friendlies, such as the U-boat tactic of diving after firing.

\textit{Homing munitions have limited autonomy}

Homing munitions are not “autonomous weapons.” It’s true that many homing
munitions are “fire and forget.” Once launched, they cannot be recalled. But this is
hardly a new development in war. Projectiles have always been “fire and forget” since
the sling and stone. Rocks, arrows, and bullets can’t be recalled after being released.
What is new – and what constitutes some limited autonomy – is that homing munitions
have rudimentary onboard intelligence that guides their behavior. They can sense the
environment (the target), determine the right course of action (which way to turn), and then act (maneuvering to hit the target). They are, in essence, a simple robot.

The autonomy given to a homing munition is narrowly bounded, however. The weapon’s intelligence is used to ensure the munition hits the target that is intended by the human operator. Homing munitions aren’t designed to search for and hunt potential targets all on their own.

Some precision-guided weapons can use some fairly sophisticated features to ensure the munition hits the intended target. The Harpoon anti-ship missile, for example, has a mode where the seeker stays off while the missile uses inertial navigation to fly a zig zag pattern towards the target. Then, at the designated location, the seeker will activate to search for the target. This allows the missile to fly past other ships in the environment that are not its target without engaging them. Thus, while the missile uses automation, it does so to strike the specific target the human intended; it is not “autonomous” for the task of deciding which target to kill. Even though a human is “out of the loop”, a fire-and-forget homing munition is a “semi-autonomous weapon” because its autonomy is constrained in which tasks it can perform.

In order for homing munitions to be effective, the human operator needs to be aware of a specific target in advance. There must be some intelligence informing the human of a particular target at a specific time and place. This intelligence could come from radars on ships or aircraft, a ping on a submarine’s sonar, satellites, or another source. But homing munitions have a very limited ability in time and space to search for targets. To launch one without a specific target would be a waste.

The U.S. Defense Department considers a homing munition just one component of a broader weapon system. The weapon system consists of a sensor to detect the enemy in the first place, a human to make a decision whether to strike, and the munition to actually destroy the target. Sometimes the weapon system is contained on a single

---


platform, such as an aircraft. In the case of an AMRAAM air-to-air missile, for example, the weapon system consists of the aircraft, radar, pilot, and AMRAAM missile. All of these elements are needed for the engagement to work. It is here, within this broader decision loop of the weapon system, that the human control resides. Even if the missile itself is “fire and forget” and has a high degree of autonomy, there still is a person “in the loop” of the weapon system, making a decision about which specific target to engage.

![Weapon System OODA Loop](image)

_A weapon system consists of the components necessary to complete an entire combat OODA loop: searching for and detecting enemy targets, deciding whether to engage them, and engaging the targets._

In other cases, components of the system may be distributed across multiple physical platforms. For example, a maritime patrol aircraft might spot an enemy ship and send the location to a nearby friendly ship, which launches a missile. Defense analysts refer to this larger, distributed system with multiple components as a _battle network_. Barry Watts described the essential role battle networks play in making precision-guided weapons effective:

Because “precision munitions” require detailed data on their intended targets or aim-points to be militarily useful—as opposed to wasteful—they require “precision information.” Indeed, the tight linkage between guided munitions and “battle networks,” whose primary reason for existence is to provide the necessary targeting information, was one of the major lessons that emerged from careful study of the US-led air campaign during Operation Desert Storm.
in 1991. … [It] is guided munitions together with the targeting networks that make these munitions “smart.” [emphasis original]¹³²

Automation is used for many engagement-related tasks in weapon systems and battle networks: finding, identifying, tracking, and prioritizing potential targets; timing when to fire; and maneuvering munitions to the target. For most weapon systems today, a human makes the decision whether to engage the target. If there is a human in the loop deciding which target(s) to engage, it is a semiautonomous weapon system.

![Semi-Autonomous Weapon System](image)

In autonomous weapon systems, the entire engagement loop—searching, detecting, deciding to engage, and engaging—is automated. (For ease of use, this thesis will frequently shorten “autonomous weapon system” to “autonomous weapon.” The terms should be treated as synonymous, with the understanding that “weapon” refers to the entire system: sensor, decision-making element, and munition.) Most weapon systems in use today are semiautonomous, but a few cross the line to autonomous weapons.

---

Human-supervised autonomous weapons: Automated defensive systems

Homing munitions significantly changed warfare. Because they can precisely target ships, bases, and vehicles, multiple homing missiles can overwhelm defenders with saturation attacks. In an era of unguided munitions, defenders could simply ride out an enemy barrage, trusting that most of the incoming rounds would miss. In an era of precision-guided weapons, however, the defender must find a way to actively intercept and defeat incoming munitions before they impact. More automation – this time for defensive purposes – was the logical response.

Human-supervised autonomous weapons are widely used to defend ships, bases, or vehicles.¹³³ Once activated, these systems will engage incoming rockets, missiles, or mortars all on their own without further human intervention. Humans are on the loop, however, supervising their operation in real time.

Supervised Autonomous Weapon System (human on the loop)

Decide to Engage Target

Search for and Detect Target

Engage Target

Once activated, supervised autonomous weapons can search for, detect, decide to engage, and engage targets all on their own, but the human can intervene, if necessary.

These human-supervised autonomous weapons are necessary for circumstances in which the speed and scale of threats could overwhelm human operators, such as saturation attacks from salvos of simultaneous incoming missiles. Automated defenses are a vital part of surviving attacks from precision-guided weapons. Examples of such systems include ship-based defenses such as the U.S. Aegis combat system and Phalanx Close-In Weapon System (CIWS); land-based systems such as the Patriot air defense system; counter-rocket, artillery, and mortar systems such as the German MANTIS; and

¹³³ See Appendix C for a complete list of systems.
active protection systems for ground vehicles, such as the Israeli Trophy or Russian Arena system.

While these systems are used for a variety of different situations – to defend ships, land bases, and ground vehicles – they operate in similar ways. Humans set the parameters of the weapon, establishing which threats the system should target and which to ignore. Depending on the system, different rules may be used for threats coming from different directions, angles, and speeds. Some systems may have multiple modes of operation, allowing human in-the-loop or on-the-loop control.

Automated defensive systems are autonomous weapons, but they have been used to-date in very narrow ways. They are used for immediate defense of human-occupied vehicles and bases. They target objects, such as missiles or rockets, not people. Humans supervise their operation in real-time and can intervene, if necessary. And the humans supervising the system are physically co-located with it, which means they can physically disable the system if it stopped responding to their commands.

**Fully autonomous weapons: Loitering munitions**

Offensive autonomous weapons are not in wide use, but there are a few select examples. Loitering munitions are a special type of munition that operates differently from homing munitions. They can loiter for extended periods of time, searching for a target over a wide area and, upon finding one, kamikaze into it to destroy the target. Because they can search for targets on their own, unlike homing munitions, loitering munitions do not require precise intelligence on enemy targets before launch. A human can launch a loitering munition into a “box” to search for enemy targets without knowledge of specific targets. Some loitering munitions keep humans in the loop via a radio connection to approve targets before engagement, making them semi-autonomous weapons.134 Some, however, are fully autonomous. The Israeli Harpy is one such weapon in use today. Once launched, the Harpy flies over a wide area searching for enemy radars and, when it finds one, kamikazes into it.135


The difference between a fully autonomous loitering munition and a semi-autonomous homing munition can be illustrated by comparing the radar-hunting Harpy with the radar-homing High-speed Anti-Radiation Missile (HARM). Both attack enemy radars, but their freedom to search for targets is massively different. The HARM has a range of 90+ kilometers and a top-speed of over 1,200 kilometers per hour, so it is only airborne for approximately four and a half minutes. Because it cannot loiter, the HARM has to be launched at a specific enemy radar in order to be useful. The Harpy can stay aloft for over two and a half hours covering up to 500 kilometers of ground. This allows the Harpy to operate independently of a broader battle network that gives the human targeting information before launch. The human who launches it decides to destroy any enemy radars within a general area in space and time, but the Harpy itself chooses the specific radar it destroys.

---


### HARM vs. Harpy

<table>
<thead>
<tr>
<th></th>
<th>Type of weapon</th>
<th>Target</th>
<th>Time to search</th>
<th>Distance</th>
<th>Degree of autonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HARM</strong></td>
<td>Homing missile</td>
<td>Radars</td>
<td>Approx. 4.5 minutes</td>
<td>90+ km</td>
<td>Semi-autonomous weapon</td>
</tr>
<tr>
<td><strong>Harpy</strong></td>
<td>Loitering munition</td>
<td>Radars</td>
<td>2.5 hours</td>
<td>500 km</td>
<td>Fully autonomous weapon</td>
</tr>
</tbody>
</table>

The Harpy is not the first loitering munition. In the 1980s, the U.S. Navy deployed a loitering anti-ship missile that could search for, select, and attack Soviet ships on its own. The Tomahawk Anti-Ship Missile (TASM) was intended to be launched over the horizon at suspected locations of Soviet ships. The TASM would then fly a search pattern over a wide area, looking for the radar signatures of Soviet ships. If the missile found one, it would destroy it.\(^\text{138}\) (Despite the name, the guidance system of the TASM was quite different from the Tomahawk Land Attack Missile (TLAM), which uses digital scene mapping to follow a pre-programmed route to its target.\(^\text{139}\)) The TASM was taken out of Navy service in the early 1990s.\(^\text{140}\) While it was never fired in anger, it has the distinction being the first operational fully autonomous weapon, a significance that was not recognized at the time.

---


Tomahawk Anti-Ship Missile Mission Profile

A typical mission for a Tomahawk Anti-Ship Missile (TASM). After being launched from a ship or submarine, the TASM would cruise to the target area. Once over the target area, it would fly a search pattern to look for targets and, if it found one, attack the target on its own. (Source: U.S. Department of Defense)

In the 1990s, the United States began development on two experimental loitering munitions: Tacit Rainbow and the Low Cost Autonomous Attack System (LOCAAS). Tacit Rainbow was intended to be a persistent anti-radiation weapon. Tacit Rainbow was directed at land-based radars, much like the Harpy.\(^{141}\) LOCAAS had an even more ambitious goal: to search for and destroy enemy tanks, which are harder targets than radars because they are not emitting in the electromagnetic spectrum.\(^{142}\) Neither Tacit Rainbow nor LOCAAS were ever deployed; both were cancelled while still in development.

These examples shine a light on a common misperception about autonomous weapons, which is the notion that intelligence is what makes a weapon “autonomous.” How intelligent a system is and which tasks it performs autonomously are different


dimensions. It is freedom, not intelligence, that defines an autonomous weapon. An autonomous weapon system is one that, once activated, is intended to search for, select, and engage targets where a human has not decided those specific targets are to be engaged. Autonomous weapons can be very simple, as the TASM was and the Harpy is today. Their simplicity limits their usefulness. Hunting harder targets in more complex environment requires more intelligence. The Harpy is the only fully autonomous weapon deployed today, but as machine intelligence advances, autonomous targeting will become feasible in a wider range of situations.

Unusual cases – Mines, encapsulated torpedo mines, and Sensor Fuzed Weapon

There are a few unusual cases – weapons that seem to blur the lines between semi-autonomous and fully autonomous weapons. Mines, encapsulated torpedo mines, and Sensor Fuzed Weapon deserve mention as weapons that use automation in special ways.

Mines are simple weapons. Placed on land or at sea, mines sit and wait for their target to approach, at which point the mine explodes. While mines are automatic devices that will detonate on their own once triggered, they have very little autonomy. Mines do not move on their own. They do not patrol. They do not hunt for targets. They simply sit and wait. (Some naval mines can drift with the current.) Additionally, mines generally have very limited methods for “deciding” whether or not to fire. Mines typically have a simple method for sensing a target and, when the threshold for the sensor is reached, the mine explodes. (Some naval mines and anti-tank mines employ a counter so that they will let the first few targets pass unharmed before detonating against a ship or vehicle later in the convoy.) Mines deserve special mention, however, because their autonomy in time is virtually unbounded. Unless specifically designed to

---

143 Mines encompass a broad range of weapons, targets, and uses. Land mines include anti-personnel mines intended to kill people and anti-tank mines intended to destroy tanks and other armored vehicles. (Many anti-tank mines are equipped with anti-handling devices that will detonate if a person tries to remove the mine, but the purpose of the mine is to destroy vehicles.) Naval mines are used to target ships and come in a variety of forms. Drift mines float on or just below the water’s surface, drifting with the current. Bottom mines lay on the sea floor. Moored mines are anchored to the sea floor via a cable and float underwater. Naval mines can be triggered via direct contact or when a ship passes close enough to trigger the mine’s acoustic, pressure, magnetic, or other sensors. Naval mines can even employ “ship counters” to allow the first few ships in a group (most likely minesweepers) to pass unharmed before detonating against a later ship. Mines are generally employed in a minefield, an area that is seeded with mines in order to slow down the enemy or to redirect them to another area. Naval mines can be delivered via ships, submarines, or aircraft. Land mines can be placed by hand or delivered via aircraft or artillery.
self-deactivate after a period of time, mines can lay in wait for years, sometimes remaining active long after a war has ended.

Encapsulated torpedo mines are a special type of naval mine that blurs the line with autonomous weapons. Rather than simply exploding once activated, encapsulated torpedo mines release a torpedo that homes in on the target.¹⁴⁴ This gives encapsulated torpedo mines a much wider area over which they can engage targets. The U.S. Mk 60 CAPTOR encapsulated torpedo mine had a published range of 8,000 yards, for example. This means CAPTOR can cover a much wider area than a regular mine, which a ship would have to pass over in order to detonate.¹⁴⁵ Even though encapsulated torpedo mines are moored in place to the seabed, their ability to launch a torpedo to chase down targets gives them a much greater degree of autonomy in space than a traditional naval mine, similar to loitering munitions. As with loitering munitions, examples of encapsulated torpedo mines are rare. The U.S. Mk 60 CAPTOR was in service for throughout the 1980s and 1990s but has since been retired. The only encapsulated torpedo mine still in service is the Russian PMK-2, which is used by Russia and China.¹⁴⁶

Sensor Fuzed Weapon (SFW) is an air-delivered anti-tank weapon that defies simple categorization. Released from an aircraft, SFW can destroy an entire column of enemy tanks within seconds. SFW functions through a series of steps: First, the aircraft releases a bomb-shaped canister than glides towards the target area. As the canister approaches the target area, the outer casing releases, exposing ten sub-munitions which are ejected from the canister. Each submunition releases a drogue parachute slowing its descent. At a certain height above the ground, the submunition springs into action. It opens its outer case, exposing four internally-held “skeets” which are then rotated out of the inner casing and exposed. The parachute releases and the submunition fires retrojets that causes it to climb in altitude while spinning furiously. The hockey-puck shaped

---


skeets are then released, flung outward violently from the force of the spinning. Each skeet carries onboard laser and infrared sensors that it uses to search for targets beneath it. Upon detecting a target (a vehicle), the skeet fires an explosively formed penetrator – a metal slug – downward into the vehicle. The metal slug strikes the vehicle on top, where vehicles have the thinnest armor, destroying it. A single Sensor Fuzed Weapon can take out a group of tanks or other armored vehicles simultaneously, with the skeets targeting each vehicle precisely.147

Similar to the distinction between Harpy and HARM, the critical variable in the evaluating SFW’s autonomy is its freedom in time and space. While the weapon distributes 40 skeets over several acres, the time the weapon can search for targets is miniscule. Each skeet can hover with its sensor active for only a few seconds before firing. Unlike the Harpy, the SFW cannot loiter for an extended period of time over hundreds of kilometers. The person launching SFW must have a reasonable degree of confidence that there is a group of tanks at a particular point in space and time to warrant launching the weapon. SFW is different than a traditional homing munition, because SFW can hit multiple objects. They must be close together; the human employing SFW launches it at a group of tanks or armored vehicles, and SFW precisely targets the vehicles individually. Thus, SFW functions like a salvo of 40 homing munitions launched at a tightly geographically clustered set of targets.

While SFW is an air-launched weapon, there have been similar ground-launched weapons that have been proposed and, in some cases, developed and used. These include the Brilliant Anti-Tank Munition (BAT), M898 Sense and Destroy Armor Munition (SADARM), SMArt 155, M93 Hornet Wide Area Munition (WAM), and the SFW submunition BLU-108/B from a surface-launched delivery system. The SADARM was reportedly used in combat in the 2003 invasion of Iraq, achieving 48 vehicle kills. Surface-placed intelligent anti-armor concepts include the Hand-Emplaced Hornet and XM1100 Scorpion.148

---


Why aren’t there more autonomous weapons?

Automation has been used extensively in weapons around the world for decades, but the amount of autonomy given to weapons has been, up to now, fairly limited. Homing munitions have seekers, but their ability to search for targets is narrowly constrained in time and space. Human-supervised autonomous weapons are only used for limited defensive purposes. Loitering munitions are offensive autonomous weapons that hunt for targets on their own, but only a handful of examples exist. The technology to build simple loitering munitions like TASM and Harpy has existed for decades. Yet, unlike homing munitions, which have widely proliferated around the globe, examples of loitering munitions are rare. Why aren’t loitering munitions in more widespread use?

We can’t know for certain why militaries have not been more aggressive in building loitering munitions, but the U.S. experience with TASM may shed some light on the question. TASM was in active service for a period of time and then was retired. According to retired Naval Officer Bryan McGrath, who is familiar with the TASM and other anti-ship missiles such as the Harpoon and was trained on the TASM in the 1980s, one concern was that the TASM could out-range the intelligence, surveillance, and reconnaissance (ISR) used to target the weapon. “We didn't have the confidence in our ISR that we thought we needed to employ the weapon confidently,” McGrath said. The TASM’s range was long enough that it could out-range the ship’s sensors. That meant that targeting had to come from another sensor, such as a helicopter or maritime patrol aircraft that detected an enemy ship. The problem, McGrath said, was a “lack of confidence in how the targeting picture would change from the time you fired the missile until you got it downrange.” Because the target could move, unless there was an “active sensor” on the target, such as a helicopter to maintain surveillance on the target.

---

the whole time, the area of uncertainty of where the target was would grow over time. As a result, “we didn’t have any confidence in it,” McGrath explained.\textsuperscript{149}

The ability of the TASM to search for targets over a wide area mitigated, to some extent, this large area of uncertainty. If the target had moved, the TASM could simply fly a search pattern looking for it. But the TASM didn’t have the ability to accurately discriminate between enemy ships and merchant vessels. As the search area widened, the risk increased that the TASM might run across a merchant ship and strike it instead. In an all-out war with the Soviet navy, that risk might be acceptable, but McGrath said that in any situations short of that, he “always doubted whether you would get the [rules of engagement]” authorization to shoot the TASM. As a result, TASM was “a weapon we just didn’t want to fire.”\textsuperscript{150}

Another factor was that if a TASM were launched and there wasn’t a valid target within the search area of the weapon, the weapon would be wasted. McGrath said, “I would be loath to launch a weapon … on scant evidence” that there was a valid target in the search area. “I would want to know that there’s something there, even if there was some kind of end-game autonomy in place.” Why? “Because the weapons cost money,” he said, “and I don’t have a lot of them. And I may have to fight tomorrow.”\textsuperscript{151}

Modern missiles can cost upwards of a million dollars apiece. As a practical matter, militaries will want to know that there is a valid enemy target in the area before launching an expensive weapon. One of the reasons militaries have not used more loitering munitions may be the fact that the advantage they bring – the ability to launch a weapon without precise targeting data in advance – may not be of much interest.

\textit{Future weapons}

The trend of creeping automation that began with Gatling’s gun will continue. Advances in AI will enable smarter weapons, which will be capable of more autonomous operation. At the same time, another facet of the information revolution is greater networking. German U-boats couldn’t control the Wren torpedo once it was launched not because they didn’t want to; they simply had no means to do so.

\textsuperscript{149} Bryan McGrath, retired naval officer, interview by author, Washington, DC, May 19, 2016.

\textsuperscript{150} Ibid.

\textsuperscript{151} Ibid.
Modern munitions are increasingly networked to allow them to be controlled or retargeted after they’ve been launched. Wire-guided munitions have existed for decades, but are only feasible for short distances. Long-range weapons are now incorporating datalinks to allow them to be controlled via radio communication, even over satellites. The Block IV Tomahawk Land Attack Missile (TLAM-E, or Tactical Tomahawk) includes a two-way satellite communications link that allows the weapon to be retargeted in flight. The Harpy 2, or Harop, has a communications link that allows it to be operated in a human-in-the-loop mode so that the human operator can directly target the weapon.

McGrath said the feature he would most desire in a future weapon wasn’t autonomy – it was a datalink. “You’ve got to talk to the missile,” he explained. “The missiles have to be part of a network.” Connecting the weapons to the network would allow you to send updates on the target while in-flight. As a result, McGrath said, the “confidence in employing that weapon would dramatically increase.”

A networked weapon is a far more valuable weapon than one on its own. By connecting a weapon to the network, the munition becomes part of a broader system and can harness sensor data from other ships, aircraft, or satellites to assist its targeting. Additionally, the human commander can keep control of the weapon while in flight, making it less likely to be wasted. One advantage to the networked Tactical Tomahawk, for example, is the ability for humans to use sensors on the weapon to do battle damage assessment (BDA) of potential targets before striking. Without the ability to perform BDA of the target, commanders might have to launch several Tomahawks at a target in order to ensure its destruction, since the first missile might not completely destroy the target. Onboard BDA allows the commander to look at the target after the first missile.


Publicly available documents are unclear on whether the Harop retains the Harpy’s ability to conduct fully autonomous anti-radar engagements as one mode of operation. Israel Aerospace Industries, developer of the Harpy and Harop, declined to comment on details of the Harpy and Harop functionality. Israel Aerospace Industries, “Harop,” http://www.iai.co.il/2013/36694-46079-EN/Business_Areas_Land.aspx.

McGrath, interview.
hits. If more strikes are needed, more missiles can be used. If not, then subsequent missiles can be diverted while in flight to secondary targets.

Everything has a countermeasure, however, and increased networking runs counter to another trend in warfare, the rise of electronic attack. The more that militaries rely on the electromagnetic spectrum for communications and sensing, the more vital it will be to win the invisible electronic war of jamming, spoofing, and deception. In future wars between advanced militaries, communications in contested environments is by no means assured. Advanced militaries have ways of communicating that are resistant to jamming, but they are limited in range and bandwidth. When communications are denied, missiles or drones will be on their own, reliant on their onboard autonomy.

Even highly advanced loitering munitions will fall into the same trap as TASM, with commanders hesitant to fire them unless targets are clearly known because of their cost. Drones change this equation, however. Drones can be launched, sent on patrol, and can return with their weapons unused if they do not find any targets. This seemingly simple feature dramatically changes how a weapon might be used. Drones could be given authority to search over a wide area in space and time to hunt for enemy targets. If there were none, the drone could return to base and be ready to search again another day.

For now, drones are used as part of traditional battle networks, with decision-making residing in the human controller. If communications links are intact, then countries can keep a human in the loop to authorize targets. If communications links are jammed, however, militaries will have to program guidance about what to do: Return home? Carry out surveillance missions? Strike fixed targets preauthorized by humans (much like cruise missiles today)? If the drones run across emerging targets of opportunity that have not been authorized in advance by a human – will they be authorized to fire? If the drones are fired upon, will they be allowed to fire back? Will they be authorized to shoot first? These are not hypothetical future questions. As robotic systems mature for operations in contested areas, nations will have to address these questions.
PART II: AUTONOMY IN WEAPONS TODAY AND FUTURE TRENDS

With each generation, military robotic systems incorporate greater autonomy. Part II examines over 25 case studies of military systems in development in order to understand how nations are incorporating autonomy into next-generation weapons and cyber systems. These case studies are chosen to represent a diverse array of applications for autonomy in weapons. The case studies illustrate the range of capabilities available to different international actors, how different states are thinking about autonomous weapons, and the degree of transparency about their weapons development. These cases show that autonomy is increasing in a range of platforms – drones, ships, ground robots, sentry guns, missiles, and cyber systems. The broad trend towards greater autonomy in both military and commercial robotics will make the technology to build autonomous weapons widely available, including to relatively unsophisticated non-state groups and even individuals. While a handful of states have pledged not to develop autonomous weapons, most leading military powers are either being opaque about their intentions or adopting a hedging strategy, leaving the option open for future autonomous weapons development. Part II will examine autonomy in four categories of systems: robotic platforms, missiles and sensors, commercially available systems, and cyber systems.

Autonomy in Robotic Platforms

The rapid proliferation of drones portends what is to come for increasingly autonomous systems. Over 90 countries have drones today as well as non-state groups such as Hamas, Hezbollah, and ISIS. Armed drone proliferation is following close behind. Over twenty nations have or are in the process of acquiring armed drones, including many that are not major military powers such as Turkmenistan, Myanmar, and Nigeria.\footnote{Fuhrmann and Horowitz, “Droning On: Explaining the Proliferation of Unmanned Aerial Vehicles.”} A number of non-state actors have also demonstrated weaponized drones for crude attacks, underscoring the accessibility of the technology.\footnote{David Reid, “A swarm of armed drones attacked a Russian military base in Syria,” CNBC.com, January 22, 2018, \url{https://www.cnbc.com/2018/01/11/swarm-of-armed-diy-drones-attacks-russian-military-base-in-syria.html}; and Elliott C. McLaughlin, “Venezuela says it has ID'd mastermind, accomplices in apparent Maduro assassination try,” CNN.com, August 7, 2018, \url{https://www.cnn.com/2018/08/06/americas/venezuela-maduro-apparent-assassination-attempt/index.html}.}
Nor is the global robotics revolution confined to the air. South Korea has deployed an armed sentry robot to its border with North Korea. Israel has used an autonomous armed ground robot, the Guardium, to patrol the Gaza border. The Guardium can maneuver around obstacles and even identify potential targets on its own, although humans are still required for lethal force. Russia has built an array of ground robots for combat applications, including deploying the armed Uran-9 robot to Syria, and has plans for a robot tank. Armed ground robots aren’t even confined to nation-states. In 2015, photographs surfaced online of Shiite militias in Iraq operating small armed ground robots.157

Armed robots are heading to sea as well. Israel has developed an armed uninhabited boat, the Protector unmanned surface vessel (USV), to patrol its coast.158 Singapore has reportedly deployed the Protector for counter-piracy missions in the Straits of Malacca.159 Even Ecuador has an armed USV, the B.A.E. ESGRUM, produced entirely indigenously. Intended to patrol Ecuadorian waterways to counter pirates, the ESGRUM carries a rifle and a rocket launcher.160

Seven nations reportedly have experimental programs intended to mature to combat drones that would operate in contested environments. These include the U.S X-47B and RQ-170, UK Taranis, China’s Sharp Sword, Russia’s Skat, France’s nEUROn, India’s Aura, and a rumored unnamed Israeli stealth drone.161 Although likely designed to operate with protected communications links to human controllers, militaries will have to decide what actions they want the drone to carry out if (and when)


communications are jammed. Restricting the drone’s rules of engagement could mean giving up valuable military advantage. Given that over a half-dozen countries already possess the fully autonomous Harpy or Harop, it isn’t a stretch to imagine them and other countries authorizing a similar level of autonomy with a recoverable drone.\footnote{Eshel, “IAI Introduces New Loitering Weapons for Anti-Radiation, Precision Strike.”}

Case studies of robotic platforms from the United Kingdom, Russia, the United States, and South Korea are included below, along with how each nation has described how they intend to use autonomy in weapons engagements.

**Taranis**

The Taranis is a next-generation uninhabited combat aerial vehicle (UCAV) similar to that being developed by other nations. BAE Systems, developer of the Taranis prototype UCAV, has perhaps given the most extensive description of how a UCAV might be used in weapons engagements. The Taranis is a demonstrator airplane, but the British military has released a description of a planned demonstration of a weapons engagement. An infographic released by BAE explains a forthcoming field test, where the Taranis “will demonstrate the ability of a UCAS [unmanned combat air system] to: fend off hostile attack; deploy weapons deep in enemy territory and relay intelligence information.”\footnote{BAE Systems, “Taranis: Looking to the Future,” http://www.baesystems.com/en/download-en/20151124120336/1434555376407.pdf.} The simulated weapons release would occur in a semi-autonomous fashion, with human control over the engagement:

1. Taranis would reach the search area via a pre-programmed flight path in the form of a three-dimensional corridor in the sky. Intelligence would be relayed to mission command.
2. When Taranis identifies a target it would be verified by mission command.
3. On the authority of mission command, Taranis would carry out a simulated firing and then return to base via the programmed flight path.

At all times, Taranis will be under the control of a highly-trained ground crew. The Mission Commander will both verify targets and authorise simulated weapons release.\footnote{Ibid.}
Thus, for this weapons engagement the human would choose the target. This vision of semi-autonomous control is consistent with other statements by BAE leadership. In a 2016 panel at the World Economic Forum in Davos, BAE Chairman Sir Roger Carr, described autonomous weapons as “very dangerous” and “fundamentally wrong.” Carr made clear that BAE only envisioned developing weapons that kept a connection to a human who could authorize and remain responsible for lethal decision-making.\textsuperscript{165} In a 2016 interview, Taranis program manager Clive Marrison similarly stated, “decisions to release a lethal mechanism will always require a human element given the Rules of Engagement used by the UK in the past.” Marrison then hedged, saying, “but the Rules of Engagement could change.”\textsuperscript{166}

The British government reacted swiftly. Following multiple media articles alleging BAE was building in the option for Taranis to “attack targets of its own accord,” if it were so authorized, the UK government released a statement the next day stating:

\begin{quote}
The UK does not possess fully autonomous weapon systems and has no intention of developing or acquiring them. The operation of our weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability for their use.\textsuperscript{167}
\end{quote}

The British government’s full-throated denial of autonomous weapons would appear to be as clear a policy statement as there could be, but an important asterisk is needed regarding how the UK defines an “autonomous weapon system.” In its official policy expressed in \textit{UK Joint Doctrine Publication 0-30.2: Unmanned Aircraft Systems}, the British military describes an autonomous system as one that is “capable of understanding higher-level intent and direction.” Short of that, a system is defined as

\begin{flushright}
\end{flushright}

\begin{flushright}
\textsuperscript{166}John Ingham, “WATCH: Unmanned Test Plane Can Seek and Destroy Heavily Defended Targets,” Express.co.uk, June 9, 2016, \url{http://www.express.co.uk/news/uk/678514/WATCH-Video-Unmanned-test-plane-Taranis}.
\end{flushright}

\begin{flushright}
\end{flushright}
“automated.” This definition of autonomy, which hinges on the complexity of the system rather than its function, is a different way of using the term autonomy than many others in discussions on autonomous weapons. The UK’s stance is not a product of sloppy language; it’s a deliberate choice. The UK doctrine publication continues:

Fully autonomous weapons systems as we describe them (machines with the ability to understand higher-level intent, being capable of deciding a course of action without depending on human oversight and control) currently do not exist and are unlikely in the near future.

This definition shifts the lexicon on autonomous weapons dramatically. When the UK government uses the term “autonomous system,” they are describing systems that are far more advanced than those existing today, not in terms of the tasks they perform, but in terms of the complexity of their intelligence. The effect of this definition is to shift the debate on autonomous weapons to far-off future systems and away from potential near-term weapon systems that may search for, select, and engage targets all on their own – what others might call “autonomous weapons.” Indeed, in its 2016 statement to the United Nations meetings on autonomous weapons, the UK stated: “The UK believes that [lethal autonomous weapon systems] do not, and may never, exist.”

The result of this position is that the UK may develop weapons that would search for, select, and engage targets on their own; it simply would call them “automated weapons,” not “autonomous weapons.” In fact, in its doctrine publication, the UK refers to systems such as the Phalanx as “automated weapon systems.”

These “automated weapon systems” may operate without a human in the loop. The UK Ministry of Defence states that automated weapon systems will always operate

---


169 Ibid, 43.


171 Ibid, 42.
under human control, but makes clear that control could be at a higher level, setting the parameters of the weapon system’s operation.

The UK Government’s policy is clear that the operation of UK weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability. Whilst weapon systems may operate in automatic modes there is always a person involved in setting appropriate parameters.\textsuperscript{172}

While this statement guarantees some human involvement, it is human involvement at a higher level of operation, setting the parameters of the system, not necessarily approving each target. In practice, therefore, the UK government’s stance on autonomous weapons retains a significant amount of flexibility. Humans will remain involved in lethal decision-making … at some level. That might mean a human operator launching an autonomous/automated weapon into an area and delegating it the authority to search for, select, and engage targets on its own. Whether the public would react differently to such a weapon if it was re-branded an “automated weapon” is unclear.

\textit{Uran-9}

In March 2018, Russia briefly deployed the Uran-9 armed ground robot to Syria. The vehicle reportedly performed poorly, but Russia continues to be one of the leaders in pushing the boundaries of armed ground robots.\textsuperscript{173} Russia is developing a fleet of ground robots for a range of missions, from protecting critical installations to urban combat and many of Russia’s robots are armed. They vary in size from the Platform-M, roughly the size of a four-wheeler armed with a grenade launcher and an assault rifle, to the much heftier Uran-9. The Uran-9 is the size of a small armored personnel carrier and has a 30mm cannon and an elevated platform to launch anti-tank guided missiles. The missiles rest on two platforms on either side of the vehicle that, when raised, allow the robot to fire missiles while safely sitting behind cover, for example behind the protective slope of a hillside. This makes the Uran-9 potentially a useful weapon in high-intensity combat against NATO forces on the plains of Europe. Uran-9s could hide behind hillsides or other protective cover and launch its missiles

\textsuperscript{172} Ibid, 43.

against enemy tanks. The Uran-9 doesn’t have the armor or guns to stand toe-to-toe against a modern tank, but because it’s uninhabited, it doesn’t have to. The Uran-9 could be a successful ambush predator. Even if firing exposed its position and led it to be taken out by opposing forces, the exchange might still be a win if it took out an enemy tank. Because there’s no one inside it and the Uran-9 is significantly smaller than a tank, and presumably less expensive, Russia could field many of them on the battlefield.\(^\text{174}\)

The Uran-9’s anti-tank mission is significant because autonomously targeting tanks is far easier than targeting people, if Russia decided to authorize autonomous engagements. With large cannons and treads, tanks are distinctive military objects not easily confused with civilian objects. In addition, militaries may be more comfortable taking risk with civilian casualties or fratricide in high-intensity state-on-state warfare, when the fate of nations is at stake. In videos of the Uran-9, human operators directly control the vehicle, but if Russia decided to pursue fully autonomous target engagement, anti-tank missions would be technologically easier than anti-personnel ones.

Russia isn’t stopping at development of the Uran-9, however. It envisions even more advanced robotic systems that could not only ambush enemy tanks, but stand with them toe-to-toe and win. Russia reportedly has plans to develop a fully robotic version of its next-generation T-14 Armata tank. The T-14 Armata, which reportedly entered production as of 2016, sports a bevy of new defensive features, including advanced armor, an active protection system to intercept incoming anti-tank missiles, and a robotic turret.\(^\text{175}\) While advanced armor and active protection systems have been added to tanks from other nations, the T-14 will be the first main battle tank to sport an uninhabited turret, which will afford the crew greater protection by sheltering them within the body of the vehicle. Making the entire tank uninhabited would be the next logical step in protection, enabling a crew to remotely control the vehicle from elsewhere. While current T-14s are human-inhabited, Russia has long term plans to develop a fully robotic version. Vyacheslav Khalitov, deputy director general of UralVagonZavod, manufacturer of the T-14 Armata has stated, “Quite possibly, future


\(^\text{175}\) Alex Lockie, “Russia Claims Its Deadly T-14 Armata Tank Is in Full Production,” Business Insider, March 17, 2016, \url{http://www.businessinsider.com/russia-claims-t14-armata-tank-is-in-production-2016-3}. 

97
wars will be waged without human involvement, that is why we have made provisions for possible robotization of Armata.” He acknowledged that achieving the goal of full roboticization would require more advanced AI that could “calculate the situation on the battlefield and, on this basis, to take the right decision”.

The Russian military has faced a difficult demographic crunch due to low birth rates in the 1990s and 2000s, leading to a nearly 20 percent personnel shortfall in the ground forces (as of 2016). Perhaps driven in part by these considerations, senior Russian military commanders have stated they intend to move towards fully robotic weapons. In a 2013 article on the future of warfare, Russian military Chief of Staff General Valery Gerasimov wrote:

Another factor influencing the essence of modern means of armed conflict is the use of modern automated complexes of military equipment and research in the area of artificial intelligence. While today we have flying drones, tomorrow’s battlefields will be filled with walking, crawling, jumping, and flying robots. In the near future it is possible a fully robotized unit will be created, capable of independently conducting military operations.

How shall we fight under such conditions? What forms and means should be used against a robotized enemy? What sort of robots do we need and how can they be developed? Already today our military minds must be thinking about these questions.

This Russian interest in pursuing fully robotic units has not escaped notice in the West. In December 2015, U.S. Deputy Secretary of Defense Robert Work mentioned

---


Gerasimov’s comments in a speech on the future of warfare. As Work has repeatedly mentioned, U.S. decisions may be shaped by those of Russia and other nations.

**Sea Hunter**

*Sea Hunter* is an entirely robotic ship, a product of the Anti-submarine warfare Continuous Trail Unmanned Vessel (ACTUV) program by the U.S. Defense Advanced Research Projects Agency (DARPA). Its mission is to track enemy submarines and it is already part of the Navy’s fleet.

There are no weapons onboard *Sea Hunter*, for now. At its christening, then-Deputy Secretary of Defense Robert Work described *Sea Hunter* as a “warship.” Weapons or no, Work said it is a “fighting ship,” part of the Navy’s future “human machine collaborative battle fleet.” At $20 million apiece, *Sea Hunter* is a fraction of the cost of a new $1.6 billion Arleigh Burke destroyer. The low price allows the Navy to purchase scores of the sub-hunting ships on the cheap. Work laid out his vision for flotilllas of the Sea Hunters roaming the seas:

You can imagine anti-submarine warfare pickets, you can imagine anti-submarine warfare wolfpacks, you can imagine mine warfare flotillas, you can imagine distributive anti-surface warfare surface action groups … We might be able to put a six pack or a four pack of missiles on them. Now imagine 50 of these distributed and operating together under the hands of a flotilla commander, and this is really something.

---


183 Work, “Remarks at the ACTUV ‘Seahunter’ Christening Ceremony.”
Like many other robotic systems, *Sea Hunter* can navigate autonomously and might someday be armed. There is no indication that DoD has any intention of authorizing autonomous weapons engagements. Nevertheless, operational circumstances could lead to situations in which it was forced to confront the question. In December 2016, China seized a small underwater robot drone the United States was operating in the South China Sea.\(^{184}\) Without humans onboard or some autonomous self-defense mechanism, *Sea Hunter* would be at the mercy of competitors who could jam its communications and attempt to seize it along with its sensitive submarine-monitoring equipment. One solution would be to arm *Sea Hunter* with autonomous lethal or non-lethal self-defense capabilities. The U.S. has not commented publicly on the idea of doing so, but much like stealth combat drones it will have to address this question one way or the other as it fields additional robotic ships like *Sea Hunter*.

**SGR-A1**

South Korea’s SGR-A1 robot is an example of the challenge in discerning how much autonomy weapon systems have. The SGR-A1 is a stationary armed sentry robot intended to defend South Korean’s border against North Korea. In 2007, when the robot was first revealed, *IEEE Spectrum* reported it had a fully autonomous mode. In an interview with *IEEE Spectrum*, Samsung principal research engineer Myung Ho Yoo stated, “the ultimate decision about shooting should be made by a human, not the robot.” But the magazine’s article made clear that Yoo’s “should” was not a requirement, and that the robot did have a fully automatic mode that was an option.\(^{185}\)

The story was picked up widely, with the SGR-A1 included as an example of a real-world autonomous weapon in articles in *The Atlantic*, BBC, NBC, *Popular Science*, and *The Verge*.\(^{186}\) The SGR-A1 made *Popular Science’s* list of “Scariest Ideas

---


In depth analytic reports by the International Committee of the Red Cross, roboticist Ron Arkin at Georgia Tech, and researchers at CalPoly working for the U.S. Navy similarly cited the SGR-A1 as fully autonomous.

In the face of this negative publicity, Samsung backpedaled, saying that in fact a human was required to be in the loop. In 2010, a spokesman for Samsung clarified that “the robots, while having the capability of automatic surveillance, cannot automatically fire at detected foreign objects or figures.” Samsung and the South Korean government have been tight-lipped about details and for good reason. The SGR-A1 is designed to defend South Korea’s border with North Korea, with which South Korea is still technically at war. Few countries on Earth face as immediate and intense a security threat as South Korea. One million North Korean soldiers and the threat of nuclear weapons loom over South Korea. In the same interview in which he asserted a human will always remain in the loop, Samsung spokesman Huh Kwang-hak asserted, “the SGR-1 can and will prevent wars.”

What are the actual specifications and design parameters for the SGR-A1? It’s essentially impossible to know without directly inspecting the robot. If South Korea is willing to delegate more autonomy to their robots than other nations, however, it wouldn’t be surprising. Defending the DMZ against North Korea is a matter of survival for South Korea. Accepting the risks of a fully autonomous sentry gun may be more than worth it for South Korea if it helps enhance deterrence against North Korea.

**Autonomy in Missiles and Sensors**

In addition to robotic platforms, autonomy is also advancing in missiles and sensors. Three case studies below illustrate the kinds of systems that nations are developing and the ways in which incremental gains in various autonomous features are changing the functionality and usefulness of these systems.

---


187 “The Scariest Ideas in Science.”
The Brimstone missile

The UK Brimstone is a next-generation missile that has come under fire from critics who have questioned whether it crosses the line to too much autonomy. The Brimstone is an aircraft-launched fire-and-forget missile intended to destroy ground vehicles or small boats. The publicly available information on Brimstone is ambiguous about the weapon’s autonomy. The manufacturer, UK-based MBDA Missile Systems, advertises an “Operator-In-The-Loop capability to meet restrictive Rules of Engagement.” Operator-in-the-loop implies a semi-autonomous weapon. On the other hand, the fact that human-in-the-loop is advertised for situations where the rules of engagement would require it implies that there might be some situations where human-in-the-loop isn’t required. Does the missile have a mode that allows operation without a human in the loop?

The manufacturer’s spec sheet on the missile reports two primary modes of operation, Single Mode and Dual Mode. Single Mode uses a semi-active laser to provide “guidance all the way to the target for stationary and other targets that can be designated by laser effectively.” Because semi-active laser guidance requires a human to “paint” the target with a laser, Single Mode is clearly semi-autonomous. The missile will go to wherever the human points the laser. Dual Mode combines the semi-active laser with a millimeter-wave (MMW) radar seeker for “fast moving and maneuvering targets and under narrow Rules of Engagement.” In Dual Mode, the semi-active laser is used to designate the target, then there is a “handoff” to the MMW seeker so the weapon can zero in on moving targets at the very end stage of the engagement. Similar to Single Mode, however, the target is designated by a human painting a laser, so the missile is still semi-autonomous, even if the MMW seeker is used at the endgame.

In both of these primary modes of operation, the missile is clearly engaging targets that have been designated by a human. However, the developer also advertises another mode of operation, “a previously-developed fire-and-forget, MMW-only mode” that can be enabled “via a software role change.” The developer explains:

This mode provides through-weather targeting, kill box-based discrimination

---

188 Markoff, “Fearing Bombs That Can Pick Whom to Kill.”
190 Ibid.
and salvo launch. It is highly effective against multi-target armor formations. Salvo-launched Brimstones self-sort based on firing order, reducing the probability of overkill for increased one-pass lethality.\textsuperscript{191}

This targeting mode would allow a human to launch a salvo of Brimstones against a group of enemy tanks, letting the missiles sort out which missiles hit which tank. According to a 2015 Popular Mechanics article by David Hambling, in this mode the Brimstone is fairly autonomous:

> It can identify, track, and lock on to vehicles autonomously. A jet can fly over a formation of enemy vehicles and release several Brimstones to find targets in a single pass. The operator sets a “kill box” for Brimstone, so it will only attack within a given area. In one demonstration, three missiles hit three target vehicles while ignoring nearby neutral vehicles.\textsuperscript{192}

On its spec sheet, the developer also describes a similar functionality against fast-moving small boats, also called fast inshore attack craft (FIAC):

> In May 2013, multiple Brimstone missiles operating in an autonomous [millimeter] wave (MMW) mode completed the world’s first single button, salvo engagement of multiple FIAC, destroying three vessels (one moving) inside a kill box, while causing no damage to nearby neutral vessels.\textsuperscript{193}

When operating in MMW-only mode, is the Brimstone an autonomous weapon? Key questions are the size of the “kill box” for the missiles and the time the missiles have to loiter and search for targets. If the kill box is small and the missiles have no ability to loiter, then the Brimstone would be similar to the Sensor Fuzed Weapon. The human operator would have to know there was a group of targets – ground vehicles or small boats – at a specific point in space and time in order for the missile to be useful. On the other hand, if the kill box was large and the missile had the ability to loiter to search the

\textsuperscript{191} Ibid.
\textsuperscript{193} MBDA, “Brimstone 2 Data Sheet November 2015.”
kill box, like a TASM or Harpy, then the operator could let the weapon search for and find the correct target. The operator could potentially launch the missile into a wide kill box where enemy operations were suspected, but with little to no information about specific targets. In those cases, the weapon could be reasonably said to be choosing its own targets, as the Harpy does. The key fact is that while the Brimstone has a reported range in excess of 20 km,\(^{194}\) it cannot loiter to search for targets. The human pilot must know the specific location of the intended targets before launch in order for the weapon to be effective, making it a semi-autonomous weapon.

The Brimstone’s most innovative feature is actually not an ability to loiter to search for targets, but rather the ability for the missiles to cooperate to ensure they don’t hit the same target. The pilot can launch a salvo of multiple Brimstones against a group of targets within a kill box and the missiles themselves “self-sort based on firing order” to hit different targets. This capability makes the Brimstone especially useful for defending against enemy swarm attacks. Iran has harassed U.S. ships with swarming small boats that could overwhelm ship defenses, causing a \textit{USS Cole}-type suicide attack. Navy helicopters armed with Brimstones would be an extremely effective defense against small boat swarms, allowing pilots to take out an entire group of enemy ships at once without having to individually target each ship.

Despite its advanced features, a human still needs to launch the Brimstone at a known group of targets. Because the weapon cannot loiter over the target area, the human would need to be reasonably certain that there were valid targets – tanks, boats, or some other target – at a certain place and time in order for the weapon to be effective. Otherwise, if the targets weren’t in the kill box when the missile turned on its seeker, then the missile would be wasted. Unlike a drone, the missile can’t return to base. The salvo launch capability allows the pilot to launch multiple missiles against a swarm of targets, and the kill box functionality allows the pilot to draw a box around a group of targets, rather than select each one individually. This makes a salvo of Brimstones similar in functionality to the Sensor Fuzed Weapon that is used to take out a column of tanks. Even though the missiles themselves self-sort which missile hits which target, the human is still deciding to attack the specific group of targets. Because the human is deciding to engage the targets, even in MMW-only mode, the Brimstone is a semi-autonomous weapon.

The line between the semi-autonomous Brimstone and a fully autonomous weapon that would choose its own targets is a thin one, however. It isn’t based on the seeker or the algorithms. The same seeker and algorithms could be used on a future weapon that could loiter over the battlespace—a missile with an upgraded engine or a drone that could patrol an area. A future weapon that patrolled a kill box, rather than entered one at a snapshot in time, would be an autonomous weapon, because the human could send the weapon to monitor the kill box without knowledge of any specific targets. It would allow the human to fire the weapon “blind” and let the weapon decide if and when to strike targets.

Even if the Brimstone doesn’t quite cross the line to an autonomous weapon, it takes one more half-step towards that line, to the point where all that is needed is a light shove to cross it. A MMW-only Brimstone could be converted into a fully autonomous weapon simply by upgrading the missile’s engine so that it could loiter for longer. Or the MMW-only algorithms and seeker could be placed a drone. Perhaps most significantly, the MMW-only mode is enabled by a software change. As autonomous technology continues to advance, more missiles around the globe will step right up to—or cross—that line.

**Collaborative Operations in Denied Environments (CODE)**

A U.S. Defense Advanced Research Projects Agency (DARPA) program takes cooperative munitions a step further, with onboard communications links that allow the munitions to autonomously coordinate their actions. According to DARPA’s description, the aim of the Collaborative Operations in Denied Environments (CODE) project is to develop “collaborative autonomy—the capability of groups of [unmanned aircraft systems] to work together under a single person’s supervisory control.” In a press release, CODE program manager Jean-Charles Ledé described the project as enabling drones to work together “just as wolves hunt in coordinated packs with minimal communication.”

Ledé clarified that the drones would remain under the supervision of a human: “multiple CODE-enabled unmanned aircraft would collaborate to find, track, identify

---


and engage targets, all under the command of a single human mission supervisor.”  

Graphics on DARPA’s website depicting how CODE might work show communications relay drones linking the drone pack back to a manned aircraft removed from the edge of the battlespace. So, in theory, a human is in/on the loop.

CODE is designed for “contested electromagnetic environments,” however, where “bandwidth limitations and communications disruptions” are likely to occur. The means that the communications link to the human-piloted aircraft might be limited or might not work at all. CODE aims to overcome these challenges by giving the drones greater intelligence and autonomy so that they can operate with minimal supervision. Cooperative behavior is central to this concept. With cooperative behavior, one person can tell a group of drones to achieve a goal and allow the drones to divvy up tasks on their own.

In CODE, the drone team finds and engages “mobile or rapidly re-locatable targets,” that is, targets whose locations cannot be specified in advance by a human operator. If there is a communications link to a human, then the human could authorize targets for engagement once CODE air vehicles find them. Communications are challenging in contested electromagnetic environments, but not impossible. U.S. fifth generation fighter aircraft use low probability of intercept / low probability of detection (LPI/LPD) methods of stealthy communicating inside enemy airspace. While these communications links are limited in range and bandwidth, they do exist. According to CODE’s technical specifications, developers should count on no more than 50 kilobits per second of communications back to the human commander, essentially the same as a 56K dial-up modem circa 1997.

Keeping a human in the loop via a connection on par with a dial-up modem would be a significant change from today, where drones stream back high-definition full

---

197 Ibid.

198 DARPA, “Collaborate Operations in Denied Environments.”

199 Ibid.


motion video. How much bandwidth is required for a human to authorize targets? Not much, in fact. The human brain is extremely good at object recognition and can recognize objects even in relatively low resolution images. Images of military objects and the surrounding area on the order of 10 to 20 kilobytes in size may be fuzzy to the human eye, but are still high enough resolution that an untrained person can discern trucks or military vehicles. A 50 kilobit per second connection could transmit one image of this size every two to three seconds (1 kilobyte = 8 kilobits). This would allow the CODE air vehicles to identify potential targets and send them back to a human supervisor who would approve (or disapprove) each target before attack.

But is this what CODE intends? CODE project documents explain that the aircraft will operate under “under a single person’s supervisory control,” but they do not specify that the human would need to approve each target before engagement. As is the case with all of the systems encountered so far, from thermostats to next-generation weapons, the key is which tasks are being performed by the human and which by the machine. Publicly available information on CODE presents a mixed picture.

A May 2016 video released online of the human-machine interface for CODE shows a human authorizing each specific individual target. The human isn’t piloting the vehicles or even directing them individually. Rather, the air vehicles are organized into four groups labeled Aces, Badger, Cobra, and Disco, each with two to four air vehicles per group. The groups are given high-level commands, such as “orbit here” or “follow this route.” Disco Group is sent on a search and destroy mission: “Disco Group search and destroy all [anti-aircraft artillery] in this area.” The human commander sketches a box with his cursor and the vehicles in Disco Group move into the box. “Disco Group conducting search and destroy at Area One,” the computer confirms.

As the air vehicles in Disco Group find suspected enemy targets, they cue up their recommended classification to the human for his confirmation. In the video, you see the human move the cursor to click “Confirm SCUD” and “Confirm AAA” [anti-aircraft artillery]. But confirmation does not mean approval to fire, at least in this video. A few seconds later, a beeping tone indicates that Disco Group has drawn up a strike plan on a target and is seeking human approval. Disco Group has 90% confidence it has found an SA-12 surface-to-air missile system and includes a photo for confirmation. The human clicks on the strike plan for more details. Beneath the picture of the SA-12 is a diagram showing estimated collateral damage. Surrounding the target is a brown splotch showing potential damage to anything in the vicinity. Just outside of the splotch is a hospital, but it is outside of the anticipated area of collateral damage. The human
clicks “Yes” to approve the engagement. In this video, a human is clearly in the loop. Many tasks are automated, but the human is exercising both high-level supervision over the mission and approval of each specific engagement.

In other public information, however, CODE seems to leave the door open to removing the human from the loop. A different video shows two teams of air vehicles, Team A and Team B, sent to engage a surface-to-air missile launcher. The specific target is identified by a human ahead of time, who then launches the missiles to take it out. The air vehicles maneuver around pop-up threats, working cooperatively sharing data while in flight. They share navigational information, allowing them to maintain a precise formation even in a GPS-denied environment. They also share sensor data, fusing information from multiple vehicles to get pictures of the target from multiple angles, allowing a higher confidence in their target identification. Then a “critical pop-up target” emerges. This isn’t their primary target, but destroying it is a high priority. Team A reprioritizes to engage the pop-up target while Team B continues to the primary target. The video makes clear this occurs under the supervision of the human commander. This implies a different type of human-machine relationship than the earlier CODE video. In this video, instead of the human in the loop, the human is on the loop, at least for pop-up threats. For their primary target, they operate in a semi-autonomous mode. The human chose the primary target. But when a pop-up threat emerges, the missiles have the authority to operate as supervised autonomous weapons. They don’t need to ask additional permission to take out the target.

DARPA’s description of CODE online seems to show a similar flexibility for whether humans would have authority for selecting and engaging targets. The CODE website says: “Using collaborative autonomy, CODE-enabled unmanned aircraft would find targets and engage them as appropriate under established rules of engagement... and adapt to dynamic situations such as ... the emergence of unanticipated threats.” This appears to leave the door open to autonomous weapons that would find and engage targets on their own.

---


204 DARPA, “Collaborative Operations in Denied Environments.”
The detailed technical description issued to developers provides additional information, but little clarity. With regards to the human-system interface, DARPA explains that developers should:

Provide a concise but comprehensive targeting chipset so the mission commander can exercise appropriate levels of human judgment over the use of force or evaluate other options.\(^{205}\)

The specific wording used, “appropriate levels of human judgment,” may sound vague, but it isn’t by accident. This guidance directly quotes the official DoD policy on autonomy in weapons, *DoD Directive 3000.09*, which states:

Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.\(^{206}\)

Notably, however, that policy does not prohibit autonomous weapons. That is, “appropriate levels of human judgment” could include autonomous weapons. In fact, the DoD policy document goes on to outline a path through which developers could seek special approval to build and deploy autonomous weapons, with appropriate safeguards and testing, should they be desired.\(^{207}\)

At a minimum, then, CODE would seem to allow for the possibility of autonomous weapons. The aim of the project is not necessarily to build autonomous weapons. The aim is to enable collaborative autonomy. But in a contested electromagnetic environment where communications links to the human supervisor might be jammed, the program appears to allow for the possibility that the drones could be delegated the authority to find and engage targets on their own.

In fact, CODE even hints at one way that collaborative autonomy might aid in target identification. Program documents list one of the advantages of collaboration as “providing multi-modal sensors and diverse observation angles to improve target


\(^{207}\) Ibid, 7-8.
identification.” Historically, automatic target recognition (ATR) algorithms have not been good enough to trust with autonomous engagements. This poor quality of ATR algorithms could be compensated for by bringing together multiple different sensors to help improve the confidence in target identification or by viewing a target from multiple angles, building a more complete picture. This is, in fact, precisely what one of the CODE videos shows, with air vehicles viewing the target from multiple directions. Whether target identification could be improved enough to allow for autonomous engagements is unclear, but if CODE is successful, U.S. defense policymakers will have to confront the question of whether to authorize autonomous weapons.

**Target Recognition and Adaption in Contested Environments (TRACE)**

While CODE aims to compensate for poor ATR algorithms by leveraging multiple sensors or multiple viewing angles of a target, DARPA’s Target Recognition and Adaption in Contested Environments (TRACE) aims to improve the target recognition algorithms directly. The project description explains the problem:

In a target-dense environment, the adversary has the advantage of using sophisticated decoys and background traffic to degrade the effectiveness of existing automatic target recognition (ATR) solutions. … the false-alarm rate of both human and machine-based radar image recognition is unacceptably high. Existing ATR algorithms also require impractically large computing resources for airborne applications.

TRACE’s aim is to overcome these problems and “develop algorithms and techniques that rapidly and accurately identify military targets using radar sensors on manned and unmanned tactical platforms.” In short, TRACE’s goal is to solve the ATR problem. This is a significant goal. Historically, ATR algorithms have fallen far short of human abilities in object identification and classification. However, recent advances in deep

---


neural networks have led to significant improvement in AI-based perception and object recognition.

Neural networks are a type of AI approach that is loosely inspired by biological neurons in animal brains. Rather than follow a script of if-then steps for how to perform a task, neural networks work based on the strength of connections within a network. The weights of various connections between nodes in the network can be changed until the correct output is achieved, such as accurately recognizing an image. In this way, neural networks “learn.” Thousands or even millions of data samples are fed into the network and network settings are constantly adjusted to “train” the network on the data until the correct output is achieved, such as the correct image category (for example, cat, lamp, car).

Deep neural networks are those that have multiple “hidden” layers between the input and output. Adding more layers in the network between the input data and output allows for a much greater complexity of the network, allowing the network to handle more complex tasks. This complexity, it turns out, is essential for image recognition, and deep neural nets have made tremendous progress. In 2015, a team of researchers from Microsoft announced that they had created a deep neural network that for the first time surpassed human performance in visual object identification.

Deep neural network has hidden layers between the input and output layers. Some deep neural networks can have as many as 150 or more hidden layers.

---


TRACE intends to harness these advances and others in machine learning to build better ATR algorithms.\textsuperscript{213} ATR algorithms that performed on par or better than humans in identifying targets such as tanks, mobile missile launchers, or artillery would be a game-changer for of finding and destroying enemy targets. If the resulting target recognition system was of sufficiently low power to be located on board the missile or drone itself, human authorization would not be required, at least from a purely technical point of view. The technology would enable weapons to hunt and destroy targets all on their own.

**Commercially-Available Components for Autonomous Weapons**

Autonomy is increasing not only in military systems designed by nation-states, but also in the crude do-it-yourself (DIY) robotic weapons built by individuals and non-state actors. In January 2018, Syrian rebels launched a mass drone attack against a Russian air base using thirteen homemade drones.\textsuperscript{214} In August 2018, two modified DJI M600 drones, retailing for $5,000 apiece, were used in a reported drone-based assassination attempt on Venezuelan President Nicolas Maduro, with each drone each carrying a kilogram of explosives.\textsuperscript{215} In the United States, hobbyists have posted videos online of themselves arming drones with handguns and even a flamethrower.\textsuperscript{216} Crude robotic weapons are already in use by non-state groups and individuals. While there

---


\textsuperscript{215} Some accounts suggest there were three drones used in the attack. Elliott C. McLaughlin, “Venezuela says it has ID'd mastermind, accomplices in apparent Maduro assassination try,” CNN.com, August 7, 2018, \url{https://www.cnn.com/2018/08/06/americas/venezuela-maduro-apparent-assassination-attempt/index.html}.

\textsuperscript{216} Rick Stella, “Update: FAA Launches Investigation into Teenager’s Gun-Wielding Drone Video,” Digital Trends, July 22, 2015, \url{http://www.digitaltrends.com/cool-tech/man-illegally-straps-handgun-to-a-drone/}. The FAA and FBI investigated but apparently there are no laws against firing weapons from a drone, provided it’s done on private property.
have yet to be examples of homemade DIY autonomous weapons that are targeting on their own, all of the components exist to build one today.

An autonomous weapon is one that can search for, select, and engage targets. That requires three abilities: the ability to maneuver intelligently through the environment to search; the ability to discriminate among potential targets to identify the correct ones; and the ability to engage targets, presumably through force. The last element has already been demonstrated – people have built homemade armed drones. The first element, the ability to autonomously navigate and search an area, is already available outdoors and is coming soon indoors. The DJI Spark, which retailed for $359 as of January 2019, can autonomously track and follow moving objects, avoid obstacles, and return home when it is low on batteries. Target identification is the only piece remaining, but all of the tools to do so are readily available for free online. Trained neural networks can be downloaded for free online that can find human faces, determine age and gender, or label human emotions.

Nor are the skills needed to use this technology difficult to come by. Courses in neural networks are available at the undergraduate (200-level) at some universities, for free online, and even at some advanced math and science magnet high schools. All of the tools to build a homemade autonomous weapon are readily available and within reach of a reasonably competent computer programmer for under $1000.

**Cyberspace and Autonomous Weapons**

Autonomy is not only increasing in physical platforms but in cyber tools as well. Automation has long played a role in malware. Worms actively transmit themselves across computer networks, a simple malicious use of automation dating

---


218 For just a few examples, see: Keras, “Using Pre-Trained Models,” https://keras.rstudio.com/articles/applications.html.


back to the 1980s. Today, automation and autonomy is increasing in both cyber offensive and defensive applications.

Stuxnet

Stuxnet, an advanced computer worm that emerged on the internet in 2010, is one example of technology pushing the boundaries of autonomy in offensive cyber weapons. Stuxnet was a form of malware that security professionals had long speculated was possible but had never before been seen: a digital weapon. It could do more than spy, steal things, or delete data. Stuxnet could break things, not just in cyberspace but in the physical world as well.

Stuxnet had a significant amount of automation. Spread via removable USB drives, the first thing Stuxnet did when it spread to a new a system was to give itself “root” access in the computer, essentially unlimited access. Then it hid, using a real security certificate from a reputable company to mask itself from anti-virus software. Then Stuxnet began searching. It spread to every machine on the network looking for a very particular type of software, Siemens Step 7, which is used to operate programmable logic controllers (PLCs) used in industrial applications. PLCs control power plants, water valves, traffic lights, and factories. They also control centrifuges in nuclear enrichment facilities.

Stuxnet wasn’t just looking for any PLC. Stuxnet operated like a homing munition, searching for a very specific type of PLC, one configured for frequency-converter drives, which are used to control centrifuge speeds. If it didn’t find its target, Stuxnet went dead and did nothing. If it did find it, then Stuxnet sprang into action, deploying two encrypted “warheads,” as computer security specialists described

---


them. One of them hijacked the PLC, changing its settings and taking control. The other recorded regular industrial operations and played them back to the humans on the other side of the PLC.

Computer security specialists widely agree that Stuxnet’s target was an industrial control facility in Iran, likely the Natanz nuclear enrichment facility. Nearly sixty percent of Stuxnet infections were in Iran and the original infections were in companies that have been tied to Iran’s nuclear enrichment program. Stuxnet infections appear to be correlated with a sharp decline in the number of centrifuges operating at Natanz. Security specialists have further speculated that the United States, Israel, or possibly both were behind Stuxnet, although attribution is difficult in cyberspace.

Stuxnet had a tremendous amount of autonomy. It was designed to operate on “air-gapped” networks, which aren’t connected to the Internet for security reasons. In order to reach inside these protected networks, Stuxnet spread via removable USB flash drives. This also meant that once Stuxnet arrived at its target, it was on its own. Computer security company Symantec described how this likely influenced Stuxnet’s design:

> While attackers could control Stuxnet with a command and control server, as mentioned previously the key computer was unlikely to have outbound Internet access. Thus, all the functionality required to sabotage a system was embedded directly in the Stuxnet executable.

---

223 Gross, “A Declaration of Cyber War.”
228 Markoff and Sanger, “In a Computer Worm, a Possible Biblical Clue.” Gross, “A Declaration of Cyber War.”
Unlike other malware, it wasn’t enough for Stuxnet to give its designers access. Stuxnet had to perform the mission autonomously.

Like other malware, Stuxnet also had the ability to replicate and propagate, infecting other computers. Stuxnet spread far beyond its original target, infecting over 100,000 computers. Symantec referred to these additional computers as “collateral damage,” an unintentional side effect of Stuxnet’s “promiscuous” spreading that allowed it to infiltrate air-gapped networks.  

To compensate for these collateral infections, however, Stuxnet had a number of safety features. First, if Stuxnet found itself on a computer that did not have the specific type of PLC it was looking for, it did nothing. Second, each copy of Stuxnet could spread via USB to only three other machines, limiting the extent of its proliferation. Finally, Stuxnet had a self-termination date. On June 24, 2012, it was designed to erase all copies of itself. (Some experts saw these safety features as further evidence that it was designed by a Western government.)

By using software to actively sabotage an industrial control system, something cyber security specialists thought was possible before Stuxnet but had not yet happened, Stuxnet was the first cyber weapon. More will inevitably follow. Stuxnet is an “open source weapon” whose code is laid bare online for other researchers to tinker with, modify, and repurpose for other attacks. The specific vulnerabilities Stuxnet exploited will have been fixed, but its design is a blueprint for cyberweapons to come.

*Autonomy in offense and defense*

Autonomy is essential to offensive cyber weapons such as Stuxnet that are intended to operate on closed networks separated from the Internet. Once it arrives at its target, Stuxnet carries out the attack on its own. In that sense, Stuxnet is analogous to a homing munition. A human chooses the target and Stuxnet conducts the attack.

Autonomy is also essential in cyber defense. The sheer volume of attacks means it is impossible to catch them all. Some will inevitably slip through defenses, whether

230 Ibid, 7.

231 Ibid,10.

232 Ibid,18.

233 Gross, “A Declaration of Cyber War.”

by using zero-day vulnerabilities, finding systems that have not yet been updated, or exploiting users who insert infected USB drives or click on nefarious links. This means that in addition to keeping malware out, security specialists have also adopted “active cyber defenses” to police networks on the inside to find malware, counter it, and patch network vulnerabilities.

Retired General Keith Alexander, former head of the National Security Agency, has argued for the necessity of automation in defending 15,000 “enclaves” (separate computer networks) within the Department of Defense. Fixing vulnerabilities at “manual speed,” he told the Senate in 2015, took months. “It should be automated,” Alexander argued. “The humans should be out of the loop.” Computer security researchers are already working to develop these more sophisticated cyber that would take humans out of the loop. As in other areas of autonomy, DARPA is at the leading edge of this research.

DARPA Cyber Grand Challenge

In 2016, DARPA hosted a Cyber Grand Challenge to advance the field of cyber security. Over one hundred teams competed to build a fully autonomous Cyber Reasoning System to defend a network. The systems competed in a live Capture the Flag competition to automatically identify computer vulnerabilities and either patch or exploit them.

David Brumley is a computer scientist at Carnegie Mellon University and CEO of ForAllSecure, whose system Mayhem won the Cyber Grand Challenge. Brumley describes his goal as building systems that “automatically check the world's software for exploitable bugs.” Mayhem realizes that vision, a “fully autonomous system for finding and fixing computer security vulnerabilities.” In that sense, Mayhem is even more ambitious than Keith Alexander’s goal of just updating software automatically.


Mayhem actually goes and finds bugs all on its own – bugs that humans are not yet aware of – then patches them.

Brumley explained that there are actually several steps in this process. The first is finding a vulnerability in a piece of software. The next step is developing either an “exploit” to take advantage of the vulnerability or a “patch” to fix it. If a vulnerability is analogous to a weak lock, then an exploit is like a custom-made key to take advantage of the lock’s weakness. A patch, on the other hand, fixes the lock.

Developing these exploits and patches isn’t enough, though. One has to know when to use them. Even on the defensive side, Brumley explained, it isn’t as simple as just applying patches as soon as they are developed. For any given vulnerability, Mayhem would develop a “suite of patches.” Fixing a vulnerability isn’t a binary thing – either it’s fixed or it isn’t. Brumley said, “There’s grades of security, and often these have different tradeoffs on performance, maybe even functionality.” Some patches might be more secure, but would cause the system to run slower. Which patch to apply depends on the system’s use. For home use, “you’d rather have it more functional rather than 100% secure,” Brumley said. A customer protecting critical systems, on the other hand, like the Department of Defense, might choose to sacrifice efficiency for better security. When to apply the patch is another factor to consider. “You don’t install a Microsoft PowerPoint update right before a big business presentation,” Brumley said.

Today, these steps are all done by people. People find the vulnerabilities, design the patches, and upload them to an automatic update server. Even the “auto-update” function on a home computer is not actually fully automatic; there’s a human-in-the-loop to click okay for the update to proceed. Every place where there is a human in the loop slows down the process of finding and patching vulnerabilities. Mayhem, on the other hand, is a “completely autonomous system for doing all those” steps. That means it isn’t just finding and patching vulnerabilities blindly. It’s also reasoning about which patch to use and when to apply it. Brumley said it’s “an autonomous system that’s taking all of those things that humans are doing, it’s automating them, and then it’s reasoning about how to use them, when to apply the patch, when to use the exploit.”

Mayhem also deploys hardening techniques on programs. Brumley described these as proactive security measures applied to a program before a vulnerability has even been discovered to make it harder to exploit, if there are vulnerabilities. And Mayhem does all of this at machine speed.

In the Cyber Grand Challenge final round, Mayhem and six other systems competed to scan each other’s software for vulnerabilities, then exploit the weaknesses in other systems while patching their own vulnerabilities. Brumley compared the
competition to seven fortresses probing each together, trying to get into locked doors. “Our goal was to come up with a skeleton key that let us in when it wasn’t supposed to.” DARPA gave points for showing a “proof of vulnerability,” essentially an exploit or “key,” to get into another system. The kind of access also mattered – full access into the system gave more points that more limited access that was only useful for stealing information.

Mayhem and its competitors pushed the state of the art forward in cyber autonomous security. “There’s nothing like that in practice, as far as the scope of autonomy,” Brumley said. “It really was looking into the future.” Mike Walker, the DARPA program manager who ran the Cyber Grand Challenge, said that the contest was the first time that automated cyber tools had moved beyond simply applying human-generated code and into the “automatic creation of knowledge.” By autonomously developing patches, they had moved beyond automated anti-virus systems that can clean up known malware to “automation of the supply chain.” Walker said, “true autonomy in the cyber domain are systems that can create their own knowledge. … It’s a pretty bright and clear line. And I think we kind of crossed it … for the first time in the cyber grand challenge.”

Walker compared the Cyber Grand Challenge to the very first chess tournaments between computers. The technology isn’t perfect. That wasn’t the point. The goal was to prove the concept to show what can be done and refine the technology over time. Brumley said Mayhem is roughly comparable to a “competent” computer security professional, someone “just fresh out of college in computer security.” Mayhem has nothing on world-class hackers.

Mayhem doesn’t need to beat the best human hackers to fundamentally change computer security, though. As the internet colonizes physical objects all around us – bringing toasters, watches, cars, thermostats and other household objects online in the “Internet of Things,” this digitization and connectivity also bring vulnerabilities. In October 2016, a botnet called Mirai hijacked everyday networked devices such as printers, routers, DVR machines, and security cameras and leveraged them for a

---

238 Mike Walker, DARPA program manager, interview by author, telephone, December 5, 2016.

239 David Brumley, Professor of Electrical and Computer Engineering, Carnegie Mellon University and CEO, ForAllSecure, interview by author, telephone, November 24, 2016.

massive DDOS attack. Brumley said most IoT devices are “ridiculously vulnerable.” There will be an estimated 20 billion IoT devices by 2020. That means there are millions of different programs, all with potential vulnerabilities. “Every program written is like a unique lock and most of those locks have never been checked to see if they’re terrible,” Brumley said. His team looked at 4,000 commercially-available internet routers and “we’ve yet to find one that’s secure,” he said. “No one’s ever bothered to check them for security.” Checking this many devices at human speed would be impossible. There are just aren’t enough computer security experts to do it. Brumley’s vision is an autonomous system to “check all these locks” and fix them.

Once you’ve uncovered a weak lock, patching it is a choice. One could just as easily make a key – an exploit – to open the lock. There’s “no difference” between the technology for offense and defense, Brumley said. They’re just different applications of the same technology. Walker agreed. “All computer security technologies are dual use,” he said.

For safety reasons, DARPA had the computers compete on an air-gapped network that was closed off from the Internet. DARPA also created a special operating system just for this contest. Even if one of the systems was plugged into the Internet, it would need to be re-engineered to search for vulnerabilities on a Windows, Linux, or Mac machine.

Brumley envisions a world where over the next decade tools like Mayhem are used to find weak locks and patch them, shoring up cyber defenses in the billions of devices online. If Brumley is right, autonomous systems like Mayhem will make computers more secure and safer ten years from now. But autonomy will keep evolving in cyberspace, with even more advanced systems beyond Mayhem yet to come.

The next evolution in autonomous cyber defense is what Brumley calls “counter-autonomy.” Mayhem targets weak locks; counter-autonomy targets the locksmith. It “leverages flaws or predictable patterns in the adversary to win.” Counter-autonomy goes beyond finding exploits, he said, it’s about “trying to find

---


vulnerabilities in the opponent’s algorithms.” Brumley compared it to playing poker: “you play the opponent.” Counter-autonomy exploits the brittleness of autonomous systems in order to defeat them.

While counter-autonomy was not part of the Cyber Grand Challenge, Brumley said they have experimented with counter-autonomy techniques that they simply didn’t use. One tool they developed embeds a hidden exploit targeting a competitor’s autonomous system into a patch. “It’s a little bit like a Trojan horse,” Brumley said. The patch “works just fine. It’s a legitimate program.” Hidden within the patch is an exploit, though, that targets one of the common tools that hackers use to analyze patches. “Anyone who tries to analyze [the patch] gets exploited,” he said. Another approach to counter-autonomy would move beyond simply finding vulnerabilities to actually creating them. This could be done in learning systems by inserting false data into the learning process. Brumley calls this the “computer equivalent to ‘the long con,’ where our systems methodically cause our adversary’s systems to ‘mis-learn’ (incorrectly learn) how to operate.”

The future of autonomy in cyberspace

Cyber weapons of the future – defensive and offensive – will incorporate greater autonomy, just the same way that more autonomy is being integrated into physical systems. In an unpublished 2016 working paper, Brumley wrote, “Make no mistake, cyber is a war between attackers and defenders, both who co-evolve as the other deploys new systems and measures. In order to win, we must act, react, and evolve faster than our adversaries.” The same competitive dynamics that exist in other domains of war exist in cyberspace and could drive increased autonomy in cyber systems in a number of ways.

One concept is automated “hacking back.” Autonomous cyber bots like Mayhem will be part of active cyber defenses, including those that use higher-level reasoning and decision-making, but these still operate within one’s own network. Some concepts for active cyber defense move beyond policing one’s own networks into going on the offense. Hacking back is when an organization responds to a cyber attack by counterattacking, gaining information about the attacker or potentially shutting down

---

244 Brumley, interview.
245 Ibid.
246 David Brumley, “Winning Cyber Battles: The Next 20 Years.”
the computers from which the attack is originating. Because many cyber attacks involve co-opting unsuspecting “zombie” computers and repurposing them for attack, hacking back can inevitably draw in third parties. Hacking back is controversial and, if done by private actors, could be illegal. As one cyber security analyst noted, “Every action accelerates.”

Automation has been used in some limited settings when hacking back. When the FBI took down the Coreflood botnet, it redirected infected botnet computers to friendly command-and-control servers, which then issued an automatic stop command to them. However, this is another example of automation being used to execute a decision made by people, which is far different than delegating the decision whether or not to hack back to an autonomous process.

Automated hacking back would consist of delegating the actual decision whether or not to go on the counteroffensive to an autonomous system. Crossing the line to automated hacking back could be very dangerous. Patrick Lin, who has written extensively on autonomy in both military and civilian applications, warned at the United Nations in 2015, “autonomous cyber weapons could automatically escalate a conflict.”

Cyberspace could be an area where automatic reactions between nation-states happen in milliseconds. The problem is if the automated reaction of one cyber system triggers an automated reaction in another. Competitive dynamics have led to instabilities in interactions between autonomous systems in other domains. In stock-trading, for example, interactions between algorithms have contributed to flash crashes. Automated hacking back could cause a flash cyber war that rapidly spirals out of

---


control. Automated hacking back is a theoretical concept, and there are no publicly known examples of it being implemented.

Autonomy could also increase in offensive cyber weapons. Stuxnet autonomously carried out its attack, but its autonomy was very narrowly scoped. Stuxnet had a number of safeguards in place to limit its spread and effects on computers that weren’t its target, as well as a self-termination date. One could envision future offensive cyber weapons that were given more free rein. Eric Messinger has argued:

… in offensive cyberwarfare, [autonomous weapon systems] may have to be deployed, because they will be integral to effective action in an environment populated by automated defenses and taking place at speeds beyond human capacities. … [The] development and deployment of offensive [autonomous weapon systems] may well be unavoidable.

It’s not clear what an offensive autonomous weapon would look like, given the varying ways in which autonomy is already used in cyberspace. A great deal of malware is already autonomous by virtue of its ability to self-replicate. The Internet Worm of 1988, for example, is an example of a runaway, self-replicating process that cannot be stopped. This is an important dimension to malware that does not have an analogy in physical weapons. Drones and robotic systems cannot self-replicate. In this sense, malware is much more similar to biological viruses and bacteria, which self-replicate and spread from host-to-host.

But there is a critical difference between digital and biological viruses. Biological viruses (and bacteria) can mutate and adapt in response to environmental conditions. They evolve. Malware, at least today, is static. Once malware is deployed, it can spread, it can hide (as Stuxnet did), but it cannot modify itself. Malware can be designed to look for updates and spread these updates among copies of itself via peer-to-peer sharing (Stuxnet did this as well), but new software updates originate with humans.

In 2008, a worm called Conficker spread through the Internet, infecting millions of computers. As computer security specialists moved to counter it, Conficker’s designers released updates, eventually fielding as many as five different variants. These

252 Velez-Green, “When ‘Killer Robots’ Declare War.”

updates allowed Conficker’s programmers to stay ahead of security specialists, upgrading the worm and closing vulnerabilities when they were detected. This made Conficker a difficult worm to defeat. At one point in time, an estimated 8 to 15 million computers worldwide were infected.²⁵⁴

Conficker used a mixture of human control and automation to stay ahead of anti-virus specialists. Conficker’s updates came from its human designers, but it used automation to get the updates clandestinely. Every day, Conficker would generate hundreds of new domain names, only one of which would link back to its human controllers with new updates. This made the traditional approach of blocking domains to isolate the worm from its controllers ineffective.²⁵⁵ As security specialists found a method to counter Conficker, a new variant would be released quickly, often within weeks.²⁵⁶ Eventually, a consortium of industry experts brought Conficker to heel, but doing so took a major effort.²⁵⁷

Conficker’s fundamental weakness was the fact that its updates could only happen at human speed. Conficker replicated autonomously and used clever automation to surreptitiously link back to its human controllers, but the contest between the hackers and security specialists was fought at human speed. Humans were the ones working to identify the worm’s weaknesses and take it down, and humans on the other side were working to adapt the worm and keep it one step ahead of anti-virus companies.

The technology that Mayhem represents could change that. What if a piece of software turned the same tools for identifying and patching vulnerabilities and applied them to itself? It could improve itself, shoring up its own defenses and resisting attack. Brumley has hypothesized about “introspective systems.” Self-adapting software that can modify itself, rather than wait on updates from its human controllers, would be a significant evolution. The result could be robust cyber defenses … or resilient malware. At the 2015 International Conference on Cyber Conflict, Alessandro Guarino hypothesized that AI-based offensive cyber weapons could “prevent and react to countermeasures,” allowing them to persist inside networks. Such an agent would be


²⁵⁶ Ibid.

“much more resilient and able to repel active measures deployed to counter it.” The tools mode possible by DARPA’s Cyber Grand Challenge enable the creation of malware that could patch its own vulnerabilities, as well as search out and exploit the vulnerabilities of other machines. This form of intelligent, adaptive malware would be significantly more virulent than worms like Mirai or Conficker and is already possible with existing technology.

**Summary**

The case studies examined in this section make clear that the technology to build simple autonomous weapons already exists today, and next-generation robotic weapons are increasing in autonomy in a variety of dimensions. Countries around the globe are investing in more autonomous robotic vehicles, missiles, and cyber weapons. Advances on the commercial market also place crude autonomous weapons within reach of non-state groups and individuals. In some cases, moving from semi-autonomous to fully autonomous weapons may take nothing more than a software upgrade or an extension of a munition’s loiter time, enabling it to search for emerging targets. Should nations decide at a later date to deploy fully autonomous weapons, they likely could do so by modifying semi-autonomous human-in-the-loop weapons that are already in their inventory. Even for less sophisticated actors, whether or not to field at least simple autonomous weapons will be a matter of choice, not capability.

---

PART III: STABILITY AND AUTONOMOUS WEAPONS

If nations were to field autonomous weapons beyond the limited examples today, how might they affect stability? Part III first explores the concept of stability; then it examines ways that autonomous weapons might affect stability. Similar to the concerns raised by Altmann and Sauer, Part III concludes that the most significant way that autonomous weapons could affect stability would be by undermining crisis stability, escalation control, and war termination if autonomous weapons led to decreased human control over conflicts. Of particular concern are unintended behaviors of autonomous weapons due to their complexity combined with the inability of contemporary machines to understand context and a lack of human control. On the other hand, there is another argument in favor of autonomous weapons leading to greater human control over the conduct of conflicts, by eliminating unreliable human warfighters and allowing political leaders to directly exercise control over their military forces through the rules programmed into autonomous weapons. Part III therefore concludes by presenting two competing hypotheses:

H1: Widespread autonomous weapons would undermine crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could initiate, escalate, and/or make it more difficult to terminate conflicts.

The alternative hypothesis is:

H2: Widespread autonomous weapons would improve crisis stability, escalation control, and war termination by increasing political leaders’ control over their military forces, reducing the risk of accidents or unauthorized escalatory actions.

Part IV will then examine over 25 case studies of complex systems in military and non-military settings to test these hypotheses.
The Concept of Stability

Stability is a broad concept covering a number of different issues: first strike stability, crisis stability, escalation control, war termination, and offense-defense balance.

First-strike stability, or first mover advantage, is the concern that there may be a significant advantage in striking first, and that this dynamic creates a perverse incentive for preemption. Thus, nations could feel compelled to launch a preemptive attack in a crisis, out of fear that if they don’t, the other side will do so first.259

Crisis stability is a related but broader concept about ensuring that any escalation in hostilities between countries is a deliberate choice on the part of their national leadership, not an accident, miscalculation, or something they feel driven to out of fear that the other might strike first. Elbridge Colby explained in Strategic Stability: Contending Interpretations, “In a stable situation, then, major war would only come about because one party truly sought it, not because of miscalculation.” There are no known instances of entirely accidental war, but Cold War strategists worried a great deal about the potential for false alarms, miscalculations, or accidents to escalate crises or precipitate conflict.260 If anything, history suggest they should have worried more. The Cold War was rife with numerous nuclear false alarms, misunderstandings, and near-use incidents that could have led to a nuclear attack. Even in conventional crises, confusion, misunderstanding enemy intentions, and the fog of war have often played a role in escalating tensions. Escalation control remains a concern even once war is underway, as most wars are fought within some constraints.

War termination is an important component of escalation control. Policymakers need to have the same degree of control over ending a war as they do – or should – over starting one.261 If policymakers do not have complete control over their forces, because attack orders cannot be recalled or communications links are severed, or if de-escalation could leave a nation vulnerable, policymakers may not be able to de-escalate a conflict even if they wanted to.

259 For some of the debates surrounding preemption, see Dan Reiter, “Exploding The Powder Keg Myth: Preemptive Wars Almost Never Happen,” The MIT Press 20, no. 2 (Fall 1995).


261 Schelling, Arms and Influence, 203-208; and Fred Ikle, Every War Must End (New York: Columbia Classics, 2005).
Offense-defense balance is a related concept to first-strike stability that pertains to whether the relative balance of military power favors the attacker or defender. An “offense dominant” warfighting regime is one where it is easier to conquer territory; a defense-dominant regime is one where it is harder to conquer territory. Machine guns, for example, favor the defense, at least tactically. It is extremely difficult to gain ground against a fortified machine gun position. In World War I, millions died in relatively static trench warfare. Tanks, on the other hand, tactically favor the offense because of their mobility. In World War II, Germany blitzkrieged across large swaths of Europe, rapidly acquiring terrain. (Offense-defense balance is subtly different from first strike stability, which is about whether there is an advantage in making the first move. In theory, a defender could make the first move, striking preemptively.)

Weapons and Stability

Different technologies or applications of technologies have sometimes improved or harmed stability, yet how specific weapons affect stability can sometimes be counterintuitive. It is not as simple as saying that “offensive” weapons are destabilizing and “defensive” weapons are stabilizing.

One of the most important weapons for ensuring nuclear stability is the ballistic missile submarine, an offensive weapon. Extremely difficult to detect and able to stay at sea for months at a time, submarines give nuclear powers an assured second-strike capability. Even if a surprise attack somehow wiped out all of a nation’s land-based nuclear missiles and bombers, the enemy could be assured that even a single surviving submarine could deliver a devastating attack. This effectively removes any first-mover advantage. The omnipresent threat of ballistic missile submarines at sea, hiding and ready to strike back, is a strong deterrent to a first strike and helps ensure stability.

In some cases, defensive weapons can be destabilizing. National missile defense shields, while nominally defensive, were seen as highly destabilizing during the Cold War because they could undermine the viability of an assured second strike deterrent. Intercepting ballistic missiles is costly and even the best missile defense shield could not hope to stop a massive overwhelming attack. However, a missile defense shield could potentially stop a very small number of missiles. This might allow

---

a country to launch a surprise nuclear first strike, wiping out most of the enemy’s nuclear missiles, and use the missile defense shield to protect against the rest. A shield made a first-strike more viable, potentially creating a first-mover advantage and undermining stability.

For other technologies, their effect on stability was more intuitive. Satellites were seen as stabilizing during the Cold War since they gave each country the ability to observe the other’s territory and confirm (or deny) whether the other had launched nuclear weapons. Attacking satellites in an attempt to blind the other nation was seen as highly provocative, since it could be a prelude to an attack (and the now-blind country could have no way of knowing if there was, in fact, an attack underway). Placing nuclear weapons in space, on the other hand, was seen as destabilizing because it could dramatically shorten the warning time available to the defender if an opponent launched a surprise attack. Not only did this make a surprise attack more feasible, but the shortened warning time meant that the defender might be more likely to respond to false alarms, undermining crisis stability.

Nations have a common interest in avoiding instability, and during the Cold War the U.S. and U.S.S.R. cooperated to avoid weapons and deployment postures that might undermine stability. After all, while each was hostile towards the other, neither side was interested in an accidental war or a crisis that might spiral out of control into nuclear war.

These measures included international treaties regulating or banning certain weapons. The Outer Space Treaty (1967) bans placing nuclear weapons in space or weapons of any kind on the Moon. The Seabed Treaty (1971) forbids placing nuclear weapons on the floor of the ocean. The Environmental Modification Convention (1977) prohibits using the environment as a weapon of war. The Anti-Ballistic Missile Treaty (1972) strictly limited the number of strategic missile defenses the Soviet Union and the United States could deploy in order to prevent the creation of robust national missile defense shields. (The United States withdrew from the ABM Treaty in


265 Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques, 1977, [https://www.state.gov/t/isn/4783.htm](https://www.state.gov/t/isn/4783.htm).
Intermediate-range nuclear missiles were also seen as particularly destabilizing, since there would be very little warning time before they hit their targets. In 1987, after seven years of negotiations, the U.S. and U.S.S.R. signed the Intermediate-Range Nuclear Forces (INF) Treaty, eliminating ground-launched ballistic and cruise missiles with a range between 500 and 5,500 kilometers. (In 2019, after several years of Russian violations, the United States withdraw from the INF Treaty.)

In other cases, there were tacit agreements between the U.S. and U.S.S.R. not to pursue certain weapons that might have been destabilizing, even though no formal treaties or agreements were ever signed. Both countries successfully demonstrated anti-satellite weapons, but neither pursued large-scale operational deployment. The Soviet Union demonstrated an anti-satellite weapon in 1970 and had an operationally-ready system by 1973, but in 1983 cancelled their program. The United States demonstrated a non-nuclear anti-satellite capability in 1985, but did not aggressively pursue an operational capability and cancelled the program in 1988. Anti-satellite weapons could undermine stability, since they could be used to aid in a blinding first strike.

Both nations developed limited numbers of neutron bombs, a “cleaner” form of nuclear bomb that would kill people with radiation but leave buildings intact. Neutron

---


267 U.S. Departme


bombs were seen as destabilizing since they could allow an attacker to wipe out a city’s population without damaging its infrastructure.\textsuperscript{271} This could make their use potentially more likely, since an attacker could use the conquered territory without fear of harmful lingering radiation.\textsuperscript{272} In the late 1970s, U.S. plans to deploy neutron bombs to Europe caused considerable controversy, forcing the United States to change course and halt the deployment.\textsuperscript{273} Other nuclear powers also have the capability to produce neutron bombs. France and China have tested them, Israel is suspected of testing one, and India has disclosed that they have the ability to build them.\textsuperscript{274} No country has ever openly pursued building large numbers of neutron bombs, however.\textsuperscript{275}

As many of these examples illustrate, “strategic stability” often refers to nuclear stability – avoiding technologies or circumstances that could incite a nuclear war. The same basic principles can be applied to thinking about strategic-level effects, such as decisions about first-mover advantage, escalation control, and war termination, that might occur below the nuclear threshold. Some technologies have clear effects on altering various aspects of stability in non-nuclear settings. Ship-launched anti-ship missiles give a significant first-mover advantage in naval warfare. The side who strikes first, by sinking some fraction of the enemy’s fleet, instantly reduces the number of

\textsuperscript{271} This is not to say that neutron bombs did not have, in theory, legitimate uses. The United States viewed neutron bombs as valuable because they could be used to defeat Soviet armored formations on allied territory without leaving residual radiation.


\textsuperscript{275} It should be noted that, in practice, any adjustable dial-a-yield nuclear weapon that can be adjusted to 10 kilotons or less could be used as a neutron bomb.
enemy missiles threatening them. This gives a decisive advantage to whoever strikes the first blow, which could be destabilizing. Low-cost effective ship-based missile defenses, such as from electromagnetic rail guns, could change this dynamic but are not yet operationally viable.

**Autonomous Weapons and Stability**

How might autonomous weapons affect stability? The defining feature of an “autonomous weapon” is the fact that decision-making for target selection and engagement is delegated to the weapon itself. Autonomous weapons could come in many forms, from large intercontinental bombers to small ground robots or undersea vehicles. They could have long ranges or short ranges, heavy payloads or light payloads. They could operate in the air, land, sea, undersea, space, or cyberspace. This makes speculating about how autonomous weapons might affect first-strike stability or offense-defense balance very challenging, since the answer may depend on the form of the weapon. Ballistic missile submarines are stabilizing in part because of their resilience. They are difficult to track and destroy, making them assets that would be very hard to take out in a surprise attack. This is a feature of the underwater environment they inhabit. Unlike in the air, where aircraft can be seen via radar, visual, or thermal signature from long distances, it is relatively easy to hide underwater. Autonomous weapons will be subject to these same physical constraints as other weapons. However, it is still possible to draw some conclusions about how general characteristics of weapons affect stability. For example, scholars have argued that mobility aids the attacker more than the defender, who doesn’t need to move. This means, in general, weapons with high mobility tend to tilt the offense-defense balance towards the offensive. For autonomous weapons, their defining characteristic is how

---


278 Glaser and Kaufmann, “What Is the Offense-Defense Balance and Can We Measure It?”.

279 Of course, mobility can also aid the defender in some situations by making targets harder to attack and defenses more resilient. Mobile missile launchers and mobile air defense systems have significantly contributed to the development of “anti-access” capabilities that increase the cost of power projection, thus making inter-state aggression less likely. This is an example of the challenge of the concept of offense-defense balance. For critiques of offense-defense balance,
target selection and engagement decisions are made. Thus, in evaluating how autonomous weapons might affect stability, they should be considered relative to semi-autonomous weapons that would have similar physical characteristics but keep a human in the loop.

It is important to separate out the potential effects of robotics and automation in general on stability and autonomous weapons in particular. Militaries around the world are investing heavily in robotics and as the robotics revolution matures, it will almost certainly alter the military balance in significant ways. Jean-Marc Rickli has suggested that robot swarms will lead to an offense-dominant regime, since swarms could be used to overwhelm defenders. More broadly, Kenneth Payne has suggested AI systems overall would favor the offense, given their increased speed and precision relative to humans and that the offense has the initiative in attack. Peter Asaro has raised concerns that robots might lower the threshold for the use of force by reducing the risk of loss of life to the attacker, a point similarly made by Payne. These outcomes are possible, but they sometimes envision a world where only the attacker has robotic weapons and not the defender, which is not realistic. Drones, which are simple robotic systems, are proliferating rapidly. Larger, more advanced militaries will always have advantages over smaller and less advanced ones, but in general technologies that are widely available like robotics tend to narrow the gap. They benefit both sides, but technologies that are easy to adopt tend to benefit weaker militaries more, somewhat leveling the playing field.

When both sides have robots, the offense-defense balance may look different. Robots, in swarms or individually, are valuable for both offense and defense. The U.S. Navy, for example, is developing swarming uninhabited boats to defend its ships. Long-endurance uninhabited aircraft could carry lasers to shoot down ballistic and


283 Office of Naval Research, “Autonomous Swarm.”
cruise missiles, a defensive application.\textsuperscript{284} It is hard to say whether military robots as a whole favor the offense or defense – it may depend on their application.\textsuperscript{285}

Even the assumption that a heavily roboticized battlefield would result in fewer casualties deserves a closer look. Robots allow warfighters to attack from greater distances, but this has been a trend in warfare for millennia, from the invention of the sling and stone to the intercontinental ballistic missile. Increased range has yet to lead to bloodless wars. As weapons increase in range, the battlefield simply expands as both sides acquire longer-range weapons. People have moved from killing each other with spears to killing each other with missiles. The violence, however, is always inflicted on people. It will always be so, because it is pain and suffering that causes the enemy to surrender.

The more salient question is how a decision to deploy fully autonomous weapons might alter the strategic balance. Specifically, their ability to operate without communications links to human operators and at superhuman speed; the impact of removing humans as a safety in crises; their effect on human decision-making in crises; and their role in deterrence. Nine mechanisms are presented below for how autonomous weapons could affect stability. Four of these mechanisms increase stability and five undermine stability.

Communications: offense-defense balance, resilience, and recallability

One advantage of fully autonomous weapons over semi-autonomous or human-supervised autonomous weapons is that they do not require a communications link back to human controllers. This makes them more resilient to communications disruption and of particular use in environments where communications are unreliable.

Communications are more likely to be challenging on the offense, when one is operating inside enemy territory and subject to jamming. For some defensive applications, one can use hard-wired cables in prepared positions that are impervious to jamming. For example, the South Korean SGR-A1 robotic sentry gun used to defend

\begin{footnotesize}
\textsuperscript{284} Scharre, “Robotics on the Battlefield Part I: Range, Persistence and Daring.”

\textsuperscript{285} Payne makes an intriguing psychological argument that AI systems will benefit the offense. He argues that the human cognitive bias of loss aversion currently benefits the defender in human-human engagements. Because machines lack this bias, their introduction will benefit the offense. If correct, this logic would suggest that the introduction of AI systems would tilt the offense-defense balance towards offense relative to an identical world without AI systems, but there is no guarantee that an AI-enabled offense would have the greater advantage over an AI-enabled defense. See Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?”, 26; and Payne, \textit{Strategy, Evolution, and War}, 178-179.
\end{footnotesize}
the North Korean DMZ is a defensive application of autonomy. There is little benefit to a fully autonomous mode, though, since South Korea could bury hardwired cables from the guns to humans controllers in fortified bunkers. Even if speed required immediate reaction (which is unlikely for anti-personnel applications), human supervision would still be possible. For some tactical defensive applications, such as the Aegis, humans are actually physically co-located with the weapon system, making communications a non-issue. This means that fully autonomous weapons without any human supervision at all are most useful on the offensive. It would be a leap to say that they would necessarily lead to an offense-dominant warfighting regime, as that would depend on a great many other factors, many of which have nothing to do with autonomy. Autonomy in general benefits both offense and defense; 30 nations already have defensive human-supervised autonomous weapons. Moving to fully autonomous weapons, however, would seem to relatively favor the offense more than defense.

With respect to first-mover advantage, full autonomy increases resilience against a surprise attack on communications networks. This is stabilizing, since it should reduce incentives for a surprise attack. If a country only had semi-autonomous weapons that required a human in the loop for each targeting decision, an adversary might be able to diminish their offensive capabilities by attacking their communications links. A great deal of military communications is done via satellites, which are relatively vulnerable. If a military can fight effectively without reliable communications, that lowers the benefit for a would-be attacker to launching a surprise attack against their satellites (or other communications nodes). Autonomous weapons, therefore, might increase stability by reducing incentives for a first strike.

The ability to continue attacking even without reliable communications poses a problem for escalation control and war termination, however. If commanders decide they wish to call off the attack, they would have no ability to recall fully autonomous weapons for the duration of the period of time that they were out of communications. Haas has noted that the lack of recallability "presents a risk of undesirable escalation and could undermine political initiatives."

This is analogous to the circumstances of the Battle of New Orleans during the War of 1812. Great Britain and the United States ended the war on December 24, 1814, but because of the limitations of communications at the time, news did not reach British and American forces until six weeks later. The

---

286 Other military systems would of course still rely on communications nodes.

Battle of New Orleans was fought after the treaty was signed but before news reached the front. Two thousand British servicemembers died after the war was technically over.

This problem of autonomous weapons is conceptually similar to mines, which linger after a conflict, except autonomous weapons would be mobile and intelligent. Because of their ability to move and perform intelligent tasks, autonomous weapons would do more than simply sit in a static location as mines do; autonomous weapons could fight entire battles. However, their mobility also introduces an automatic self-limitation. Once they’ve exhausted their fuel reserves, autonomous weapons would crash or, depending on how they were designed, switch to a static energy-conserving mode. There are additional fail-safe measures that nations could introduce, such as self-deactivation after a set time period.

**Speed and crisis stability**

While the ability to carry out attacks without communications has a mixed effect on stability, autonomous weapons’ advantage in speed is decidedly negative for crisis stability. Autonomous weapons risk accelerating the pace of battle and shortening time for human decision-making overall. This heightens instability in a crisis. Thomas Schelling wrote in *Arms and Influence*:

The premium on haste—the advantage, in case of war, in being the one to launch it or in being a quick second in retaliation if the other side gets off the first blow—is undoubtedly the greatest piece of mischief that can be introduced into military forces, and the greatest source of danger that peace will explode into all out war.\(^{288}\)

Crises are rife with uncertainty and potential for miscalculation, and time is a virtue in such situations. Schelling explained:

it is hard to imagine how anybody would be precipitated into full-scale war by accident, false alarm, mischief, or momentary panic, if it were not for such urgency to get in quick. … But when speed is critical, the victim of an accident or false-alarm is under terrible pressure\(^{289}\)

\(^{288}\) Schelling, *Arms and Influence*, 227.

\(^{289}\) Ibid, 227.
Some forms of autonomy could help to reduce these time pressures. Semi-autonomous weapons that automated tasks like finding and identifying targets could be stabilizing, since they could buy more time for human decision-makers. Fully autonomous and supervised autonomous weapons short-circuit human decision-making, however, speeding up engagements. This could accelerate reactions and counter-reactions, making it harder for humans to understand and control events. This could be harmful even without the kinds of accidents and unexpected interactions between autonomous systems that could cause events to spiral out of control. Even if everything functions properly, policymakers could nevertheless effectively lose the ability to control escalation as the speed of action on the battlefield begins to eclipse their speed of decision-making.⁹⁰

Removing the human fail-safe

In the midst of this fast-paced environment, autonomous weapons would remove a vital safety in preventing unwanted escalation: human judgment. Petrov’s fateful decision in 1983 represents an extreme case of the benefits of human judgment, but there are many more in crisis situations. Schelling wrote about the virtues of:

- restraining devices for weapons, men, and decision-processes – delaying mechanisms, safety devices, double-check and consultation procedures,
- conservative rules for responding to alarms and communication failure, and in general both institutions and mechanisms for avoiding an unauthorized firing or a hasty reaction to untoward events⁹¹

Used in the right way, automation can act as an additional safety, such as automatic braking on cars. Automation increases safety when it is additive to human judgment, however, not a replacement for it.⁹² In semi-autonomous weapons, automation can be used to aid in human decision-making and increase precision in execution, but humans

---


⁹¹ Schelling, Arms and Influence, 231.

⁹² This doesn’t mean that automatic safeties would never take over control from humans. Often they do. Automatic braking takes over control of the car from the human driver. Similarly, Auto-GCAS on fighter airplanes take over control from the pilot. But they do so in narrow situations where precision is what is required, not judgment.
retain final decision authority. Pushing humans out of the loop, whether for speed or because of a lack of communications, removes an important safety. Autonomous weapons are the opposite of the kind of safeties that are vital to avoiding haste and unwanted escalation in a crisis.

**Command-and-control and the psychology of crisis decision-making**

Autonomous weapons would also complicate the psychology of crisis decision-making in hard-to-predict ways. Haas has argued that autonomous systems “would become an additional actor participating in the crisis, though one who is tightly constrained by a set of algorithms and mission objectives.”293 Payne has similarly explored how AI will interact with human psychology to change strategy.294

Stability is as much about perceptions and human psychology as it is the weapons themselves. Two gunslingers staring each other down aren’t trying to evaluate the caliber of the other’s gun. They’re weighing what’s in the mind of the other fighter. Machines today are woefully unequipped to perform this kind of task. Machines may outperform humans in speed and precision, but current AI cannot perform theory-of-mind tasks such as imagining another person’s intent. At the tactical level of war, this may not be important. Once the gunslinger has made a decision to draw his weapon, automating the tasks of drawing, aiming, and firing would undoubtedly be faster than doing it by hand. Likewise, once humans have directed that an attack should occur, autonomous weapons may be more effective than humans in carrying out the attack. Crises, however, are periods of militarized tension between nations that have the potential to escalate into full-blow war, but where nations have not yet made the decision to escalate. Countries are attempting to communicate their resolve – their willingness to go to war if need be – but without actually starting a war. This is a delicate balancing act. Unlike a battle, which is fought for tactical or operational advantage, crises are ultimately a form of communication between national leaders, whose intentions are communicated through military actions.

Command-and-control is a common concern in crises. National leaders do not have perfect control over their forces, and warfighters can and sometimes do take actions inconsistent with their national leadership’s intent, whether out of ignorance, negligence, or deliberate attempts to defy authorities. The Cuban missile crisis in 1962

---


was rife with such incidents. On October 26th, ten days into the crisis, authorities at Vandenberg Air Force base carried out a scheduled test launch of an Atlas ICBM without first checking with the White House.  

The next morning on October 27th, an American U-2 surveillance plane was shot down while flying over Cuba, despite orders by Soviet Premier Nikita Khrushchev not to fire on U.S. surveillance aircraft. (The missile appears to have been fired by Cuban forces on Fidel Castro’s orders.) Later that same day, another U-2 flying over the Arctic Circle accidentally strayed into Soviet territory. Soviet and American leaders could not know for certain whether these incidents were intentional signals by the adversary to escalate or individual units acting on their own. Events like these have the potential to ratchet up tensions through inadvertent or accidental escalation.

In theory, autonomous weapons ought to be the perfect soldier, precisely carrying out orders without any deviation. This might eliminate some incidents. For example, on October 24, 1962 when U.S. Strategic Air Command (SAC) was ordered to DEFCON 2, just one step short of nuclear war, SAC commander General Thomas Power openly broadcast a message to his troops on an unencrypted radio channel revealing U.S. heightened readiness levels. The open broadcast was not authorized. Unlike people, autonomous weapons would be incapable of violating their programming. On the other hand, their brittleness and inability to understand the context for their actions would be a major liability in other ways. The Vandenberg IBCM test, for example, was caused by officers following pre-established guidance without pausing to ask whether that guidance still applied in light of new information (the unfolding crisis over Cuba).

Often, the correct decision in any given moment depends not on rigid adherence to guidance, but rather on understanding the intent behind the guidance. Militaries have a concept of “commander’s intent,” a succinct statement given by commanders to


297 Powers’ motivations for sending this message have been debated by historians. See Sagan, The Limits of Safety, 68–69. Some sources incorrectly state that Powers unilaterally decided to move to DEFCON 2 without authorization, but that does not appear to be the case.

298 See also Payne, “AI, warbot”; and Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?”.
subordinates describing the commander’s goals. Sometimes, meeting the commander’s intent requires deviating from the plan because of new facts on the ground. Humans are not perfect, but they are capable of using common sense and judgment to comply with the intent behind a rule, rather than the rule itself. Humans can disobey rules and in tense situations, counterintuitively, that may be a good thing.

At the heart of the matter is whether more flexibility for how subordinates carry out directions is a good thing or a bad thing. On the battlefield, greater flexibility is generally preferred, within broad bounds of the law and rules of engagement. In “Communicating Intent and Imparting Presence,” Lieutenant Colonel Lawrence Shattuck wrote:

If … the enemy commander has 10 possible courses of action, but the friendly commander, restricted by the senior commander, still has only one course of action available, the enemy clearly has the advantage.

In crises tighter control over one’s forces is generally preferred, since even small actions can have strategic consequences. Zero flexibility for subordinates with no opportunity to exercise common sense is sure to invite disaster, though. Partly, this is because national leaders cannot possibly foresee all eventualities. War is characterized by uncertainty. Unanticipated circumstances will arise. War is also competitive. The enemy will almost certainly attempt to exploit rigid behavioral rules for their own advantage. Haas outlines a range of actions that nations could take to manipulate autonomous weapons’ rigidity in an attempt to make them attack the wrong targets or miss correct ones. Humans might be able see these ruses or deceptions for what they are and innovate on the fly, in accordance with their understanding of commander’s intent. Autonomous weapons would follow their programming. On a purely tactical level, other benefits to autonomous weapons may outweigh this vulnerability, but in crisis situations, when a single misplaced shot could ratchet up tensions towards war, careful judgment is needed.

In recent years, the U.S. military has begun to worry about the problem of the “strategic corporal.” The basic idea is that a relatively low-ranking individual could,

---

299 Headquarters, Department of the Army, Field Manual 100-5 (June 1993), 6-6.


through his or her actions on the battlefield, have strategic effects that determine the course of the war. The solution to this problem is to better educate junior leaders on the strategic consequences of their actions in order to improve their decision-making overall, rather than giving them a strict set of rules to follow. Any set of rules followed blindly and without regard to the commander’s intent can be manipulated by a clever enemy. Autonomous weapons would do precisely what they are told, regardless of how dumb or ill-conceived the orders appear in the moment. Their rigidity might seem appealing from a command-and-control standpoint, but the result is an amplification of the strategic corporal problem.

Fundamentally, though, there is a deeper problem with leaders attempting to increase their command-and-control in crises by directly programming engagement rules into autonomous weapons: leaders themselves may not be able to accurately predict what decisions they would want to take in the future. This is due to a cognitive bias known as “projection bias,” which is when humans incorrectly project their current beliefs and desires onto others and even their future selves.

David Danks argued that projection bias is “a very real problem” for autonomous weapons. Even if we could ensure that the autonomous weapon would flawlessly carry out political leaders’ directions, with no malfunctions or manipulation by the enemy, “you still have the problem that that’s a snapshot of the preferences and desires at that moment in time,” Danks said. Human preferences and desires change. Kenneth Payne captured the essence of this problem in “Artificial Intelligence: A Revolution in Strategic Affairs?”:

To meet human expectations, AI will need to faithfully reflect the constantly evolving human perception of conflict. But the question remains whether it should act on what humans wanted at the outset, what they want in the present moment, or what the AI understands to be most closely suited to the future they might want. There is scope for considerable differences among all three – with profound implications for whom the AI targets, and how.

---


Danks argued that people generally do a good job of predicting their own future preferences for situations they have experience with, but for “a completely novel situation … there’s real risks that we’re going to have pretty significant projection biases.”

There is a real world example of projection bias from the Cuban missile crisis. Robert McNamara, who was Secretary of Defense at the time, later explained that the President’s senior advisors believed that if the U-2 they sent to fly over Cuba was shot down, that would signal a deliberate move by the Soviets to escalate. They had decided ahead of time, therefore, that if the U-2 was shot down, the U.S. would attack:

[B]efore we sent the U-2 out, we agreed that if it was shot down we wouldn't meet, we'd simply attack. It was shot down on Friday. ... Fortunately, we changed our mind, we thought "Well, it might have been an accident, we won't attack."

When actually faced with the decision, McNamara and others had a different view. They did not accurately predict their own preferences for what they wanted to do after the plane were shot down. In the case of the decision to attack Cuba, McNamara and others could reverse course. They had not actually delegated the authority to attack. There was another moment during the Cuban missile crisis, however, when Soviet leaders had delegated release authority for nuclear weapons and the world came chillingly close to nuclear war.

On October 27th, the same day that the U-2 was shot down over Cuba and another U-2 flying over the Arctic strayed into Soviet territory, U.S. ships at the quarantine (blockade) line began dropping signaling depth charges on the Soviet submarine B-59 to compel it to surface. The U.S. Navy was not aware that the B-59 was armed with a nuclear-tipped torpedo with a 15 kiloton warhead, about the size of the bomb dropped on Hiroshima. Furthermore, Soviet command had delegated authority to the ship’s captain to use the torpedo if the ship was “hulled” (a hole blown in the hull from depth charges). Normally, authorization was required from two people to fire a nuclear torpedo: the ship’s captain and political officer. According to Soviet sailors

305 Danks, interview.

306 Robert McNamara, interview included as special feature on Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (DVD). Columbia Tristar Home Entertainment, (2004) [1964].
aboard the submarine, the submarine’s captain, Valentin Savitsky, ordered the nuclear torpedo prepared for launch, declaring, “We’re going to blast them now! We will die, but we will sink them all.” Fortunately, on this submarine the flotilla commander, Captain Vasili Arkhipov, was also present. As flotilla commander, he was Captain Savitsky’s superior and his approval was also required. Reportedly, only Arkhipov was opposed to launching the torpedo.307 Like Stanislav Petrov, once again the judgment of a single Soviet officer may have again prevented the outbreak of nuclear war.

Autonomous weapons would remove the capacity of humans to make judgments about escalation based on context and could lock military forces into actions that leaders find undesirable, undermining stability.

*Deterrence and the dead hand*

Sometimes, there is a benefit to tying one’s hands in a crisis. Strategists have often compared crises to a game of chicken between two drivers.308 Neither side wants a collision, but neither wants to be the first to swerve. One way to win is to demonstrably tie one’s hands so that one cannot swerve. Herman Kahn gave the example of a driver who “takes the steering wheel and throws it out the window.”309 The onus is now on the other side to swerve to avoid a collision.

Here, autonomous weapons’ rigid adherence to rules and lack of recallability would be a benefit. Once could build a robot car that would never swerve, never flinch, never even blink at a collision. It would be the perfect driver to win at chicken. Michael Horowitz has presented the thought experiment of an alternative Cuban missile crisis where the U.S. ships conducting the blockade were autonomous weapons. They would be programmed to fire on any Soviet ships crossing the blockade line. If this could be credibly communicated to Soviets, it would put the onus for avoiding conflict on the Soviets. The problem, Horowitz asked, was “how would the Kennedy Administration

---


have persuaded the Soviet Union that that was the case?” There would be no way to convincingly prove to Soviet leadership that the robotic vessels were actually programmed to fire. U.S. leaders could claim that was the case, but the claim would be meaningless, since that’s also what they would say if they were bluffing. The United States would certainly not allow the Soviets to randomly stop and inspect the code of U.S. ships at the blockade. There would be no credible way to demonstrate that one had, in fact, tied one’s hands. It would be the equivalent of ripping out the steering wheel, but being unable to throw it out the window. (Similarly, the Soviets could program their ships to run the blockade without any option for turning back, but there would be no way to prove to the Americans they had done so.)

After the Cold War, reports emerged that the Soviet Union had built a semi-automatic “doomsday device,” nicknamed “Dead Hand.” Officially called, “Perimeter,” the system was reportedly an semi-automated nuclear command-and-control system designed to allow a massive retaliatory attack in the event that a U.S. first strike took out Soviet leadership. Accounts of Perimeter’s functionality differ, but the essential idea was that the system would remain inactive during peacetime but, in the event of a crisis, it could be activated as a “fail-deadly” mechanism for ensuring retaliation. When active, a network of light, radiation, seismic, and pressure sensors would evaluate whether there had been any nuclear detonations on Soviet soil. If a nuclear detonation was detected, then the system would check for


312 Some accounts by former Soviet officials state that the Dead Hand was investigated and possibly even developed, but never deployed operationally. Andrian A. Danilevich, interview by John G. Hines, March 5, 1990, [http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Danilevich.pdf](http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Danilevich.pdf), 62-63; and Viktor M. Surikov, interview by John G. Hines, September 11, 1993, [http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Surikov.PDF](http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Surikov.PDF), 134-135. It is unclear, though, whether this refers in reference or not to a fully automatic system. Multiple sources confirm the system was active, although the degree of automation is ambiguous in their accounts: Kataev, 100-101; and Korobushin, 107.

313 Korobushin, 107; Thompson, “Inside the Apocalyptic Soviet Doomsday Machine.”

314 Ibid.
communications to the Soviet military General Staff. If communications were active, it would wait a predetermined amount of time, ranging from on the order of 15 minutes to an hour, for an order to cancel the launch. If there was no order to stop the launch, Perimeter would act like a “dead man’s switch” and would bypass normal layers of command and transfer launch authority directly to individuals within a deep underground protected bunker. There would still be a human in the loop, but the decision would reside with whichever staff officer was on duty at the time. Soviet leadership would be cut out of the loop. With a push of a button, that individual could launch a series of communications rockets that would fly over Soviet territory and beam down the nuclear launch codes to missiles in hardened silos. Soviet ICBMs would then launch a massive strike on the United States, the retaliation of a dying nation.

There was a purpose to the madness. In theory, if everything worked properly, a system like Perimeter would enhance stability. Because a retaliatory strike would be assured, the system would remove the need for haste from Soviet leaders’ decision-making in a crisis. If there were warnings of an incoming U.S. surprise attack, as was the case in the 1983 Petrov incident and again in 1995 when Russian military leaders brought the nuclear suitcase to Boris Yeltsin in response to a Norwegian scientific rocket launch, there would be no rush to respond. Soviet or Russian leaders would be in no rush to fire their nuclear missiles in an ambiguous situation, because even if the U.S. succeeded in a decapitating strike, retaliation was assured. The knowledge of this would also presumably deter the U.S. from even considering a preemptive first strike (if the Soviets told the Americans). The problem, of course, is that such a system comes with tremendous risks. If Perimeter were to falsely detect an event when activated, as the Soviet Oko satellite system did in 1983 when it falsely detected U.S. ICBM launches, or if Soviet leaders were unable to stop the mechanism once it was activated, the system would obliterate humanity.

Autonomous weapons give leaders the ability to irrevocably commit to an action, but without the desired deterrent effect because of the inability to credibly signal that commitment to an adversary. Pre-commitment could, in principle, be beneficial in changing leaders’ own calculus in a crisis, such as a “dead hand” assured retaliatory mechanism, but would run the risk of unintended outcomes from accidents or projection bias.

---

315 Ibid.
316 Kataev. Korobushin.
**Stability-instability paradox and the mad robot theory**

The logic of mutual assured destruction (MAD) is to make any nuclear attack inherently suicidal. If a retaliatory response is assured, then attacking the enemy is akin to attacking oneself. This dynamic is stabilizing because it deters both sides from using nuclear weapons. Ironically, though, strategists have worried that too much stability could be a bad thing. This is known as the “stability-instability paradox.”

The concern is that if there were too much mutual restraint surrounding nuclear weapons, then they could lose their value as a deterrent. This could embolden aggression below the nuclear threshold, since countries could be confident that an adversary would not respond with nuclear weapons. Under this logic, some instability – some risk of accidents and miscalculation – is a good thing, because it induces caution. One response to this paradox is the “madman theory.” The principle behind the madman theory, espoused by President Richard Nixon, is to convince the enemy’s leadership that a nation’s leaders are so volatile and irrational that they just might launch a nuclear attack, consequences be damned.

This suggests another way autonomous weapons might improve stability: the “mad robot theory.” If countries perceive autonomous weapons as dangerous and an unpredictable element that cannot be completely controlled, then introducing them into a crisis might induce caution. It would be the equivalent of what Schelling has described as “the threat that leaves something to chance.” By deploying autonomous weapons into a tense environment, a country would effectively be saying to the enemy, “Things are now out of my hands. Circumstances may lead to war; they may not. I cannot control it, and your only course of action if you wish to avoid war is to back down.” Unlike the problem of credibly tying one’s hands by locking in escalatory rules of engagement, this threat does not require convincing the enemy what the autonomous weapons’ rules of engagement are. In fact, uncertainty makes the “mad robot” threat more credible, since deterrence hinges on the robot’s unpredictability, rather than the

---


certainty of its actions. Deploying an untested and unverified autonomous weapon would be even more of a deterrent, since one could convincingly say that its behavior was truly unpredictable.

What is interesting about this idea is that its efficacy rests solely on humans’ perception of autonomous weapons, and not about the actual functionality of the weapons themselves. The weapons may be reliable or unreliable – it doesn’t matter. What matters is that they are perceived as unpredictable and, as a result, the other side becomes more cautious. Of course, the actual functionality of the weapons does matter when it comes to how a crisis unfolds. If the weapons are perceived as a little dangerous, so leaders become slightly more cautious than they would be otherwise, but in fact the weapons are far more dangerous than anyone believed, then the result could still be disaster. The key variable is the difference between how humans perceive the risk of autonomous weapons and their actual risk. If military and civilian leaders overestimate the risks of autonomous weapons – they perceive them to be more dangerous than they really are – then there is nothing to worry about. Their introduction into crises will induce caution but, in fact, they will be unlikely to cause harm.

How accurately people assess the risks of autonomous weapons and whether they are likely to over- or underestimate their risks hinges on both individual psychology and how organizations evaluate risk in complex systems. Danks said, “There’s a real problem here for autonomous weapons.” For starters, he explained that people are poor predictors of behavior in systems that have feedback loops, where one action creates a counter-reaction and so forth. He said that autonomous weapons interacting with the environment was an example of a feedback loop, and our understanding of human psychology suggests that people do a poor job of predicting how those situations might unfold. (Real-world experience with complex autonomous systems in uncontrolled environments, such as stock trading, lends credence to this theory.)

Furthermore, Danks said, people are poor predictors of risk for situations which they have no experience. He said that people actually are very good at estimating risks of situations they have experience with. For example, Danks explained that people are good estimators of their likelihood of getting into an automobile accident. But when they have no prior knowledge, then their ability to accurately assess risks falls apart. “Autonomous weapon systems are very new. They aren’t just a bigger gun,” he said. “If

320 Danks, interview.
you think of them as a bigger gun, you say, ‘Well we’ve got a lot of experience with guns.’” That might lead one to think that one could accurately evaluate the risks of autonomous weapons. But Danks said he thought they were “something qualitatively different” than other weapons.\textsuperscript{321}

If there is something different about autonomous weapons, then it isn’t simply a matter of doing sufficient testing to verify their functionality. How much testing is required to ensure an autonomous weapon fails less than 0.0001\% of the time, for example? We don’t know how it’s going to interact with the environment, and we can’t know until we build up experience with more sophisticated autonomous systems over time. Danks said it would be different if we already had extensive experience with safely operating complex autonomous systems in real-world environments. Unfortunately, human experience with complex systems suggests that surprise accidents are often lurking below the surface. Danks’ concluded that “it’s just completely unreasonable and hopelessly optimistic to think that we would be good at estimating the risks.”\textsuperscript{322}

**Summary**

Nine mechanisms have been presented, four of which would increase stability and five of which would decrease stability. These mechanisms are summarized in the table below, along with a brief summary of the significance of each mechanism.

<table>
<thead>
<tr>
<th>Feature of autonomous weapons</th>
<th>Hypothesized effect on stability</th>
<th>Type of stability affected</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to operate without communications</td>
<td>Reduces incentives for a first-strike against communications nodes, such as</td>
<td>First-strike stability</td>
<td>Other military forces would still rely on communications nodes, so autonomous weapons would add some measures</td>
</tr>
</tbody>
</table>

\textsuperscript{321}Danks, interview.

\textsuperscript{322}Danks, interview.
<table>
<thead>
<tr>
<th>Unable to deviate from programming</th>
<th>Will precisely follow guidance from political leaders, reducing the risk of unauthorized escalation by subordinates.</th>
<th>Crisis stability</th>
<th>Significance depends on the reliability of fully autonomous weapons and their ability to successfully carry out actions in real-world environments consistent with political leaders’ intentions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unable to deviate from programming and able to operate without communications</td>
<td>Increase deterrence by giving political leaders the ability to irrevocably commit to an action.</td>
<td>Deterrence</td>
<td>Does not appear to be a credible way for leaders to signal to others that they have committed to an action, given that the decision-making is embedded in software.</td>
</tr>
<tr>
<td>Perceived as unpredictable</td>
<td>Introducing autonomous weapons into a crisis will induce caution in policymakers, since they are unpredictable.</td>
<td>Madman theory; deterrence</td>
<td>For this mechanism to be stabilizing, policymakers must over-estimate the risks of autonomous weapons. Otherwise, the net effect of introducing autonomous weapons will be to add risk and be destabilizing.</td>
</tr>
</tbody>
</table>
### Potentially Destabilizing Effects of Autonomous Weapons

<table>
<thead>
<tr>
<th>Feature of autonomous weapons</th>
<th>Hypothesized effect on stability</th>
<th>Type of stability affected</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to operate without communications</td>
<td>Benefits the offense more than defense, because attacker likely to operate in communications-contested environments while defender may have the ability to use more resilient communications measures.</td>
<td>Offense-defense balance</td>
<td>Relatively more advantageous to attacker but does not necessarily translate to an offense-dominant regime. Overall offense-defense balance hinges on many other factors.</td>
</tr>
<tr>
<td>Able to operate without communications</td>
<td>Reduce human control over war termination because no ability to recall fully autonomous weapons for the duration of the period of time that they were out of communications.</td>
<td>War termination</td>
<td>Could be a complicating factor in some scenarios. However, risk should be manageable if militaries attend to it as militaries should have a clear idea of how long a weapon system can operate before running out of fuel. Possible to anticipate and account for risk. Similar to mines, so not a novel problem.</td>
</tr>
<tr>
<td>Able to react faster than humans</td>
<td>Accelerate the tempo of actions in a conflict or crisis, reducing the time available for humans to make informed, rational decisions about escalation.</td>
<td>Crisis stability</td>
<td>Numerous Cold War nuclear false alarms – and the ways that humans identified them as false – lend support to the idea that shortening time for human judgment in a crisis would be destabilizing.</td>
</tr>
<tr>
<td>Removes human fail-safe</td>
<td>Increase propensity for accidents by removing the role of the human operator as a potential fail-safe.</td>
<td>Escalation control</td>
<td>Significance depends on the likelihood of accidents with fully autonomous weapons and the value, or lack thereof, a human would have had in preventing those accidents.</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Unable to deviate from programming</td>
<td>Rigid adherence and inability to deviate from programming based on “commander’s intent” could lead to brittle behavior in crises, leading to unintended escalation.</td>
<td>Escalation control</td>
<td>Significance depends on the degree to which flexibility is needed to successfully execute actions in real-world environments consistent with political leaders’ intentions.</td>
</tr>
</tbody>
</table>

Some of the mechanisms would appear to have a marginal effect relative to other factors. For example, fully autonomous weapons that could operate without communications links would appear to slightly benefit the offense more than the defense, but this overall effect is likely marginal compared to many other factors that could affect offense-defense balance. A slight benefit to the offense would not necessarily equate to an offense-dominant regime. Similarly, the ability to operate without communications would lower incentives for an opponent to launch a first-strike against vulnerable communications nodes, such as satellites, but other military forces would still rely on communications. This effect would not eliminate all incentives for a first-strike and other factors could still lead to a first-strike dominant warfighting regime. The theoretical increased deterrent benefit of fully autonomous weapons does not appear executable in practice, given the impossibility of credibly demonstrating to an adversary the rules embedded in the weapon’s software.

Other attributes of autonomous weapons appear to have clear effects on stability but may be manageable by policymakers. The ability to operate without communications has many operational benefits but has a clearly deleterious effect on war termination. If a nation’s political leadership desired an immediate cease-fire, it would not be able to recall any fully autonomous weapons that were out of
communications. These weapons would continue engaging targets until they completed their mission, exhausted their ammunition or fuel reserves, or were destroyed. This risk would be entirely foreseeable and calculable by policymakers, however. Assuming autonomous weapons functioned as intended, policymakers could account for these risks and mitigate them when necessary through a variety of control measures. These include: programming that directed autonomous weapons to cease fire or return to base after a set period of time; limiting ammunition and/or fuel; and/or software- and hardware-level self-deactivation functions. This problem is also not entirely novel, so there are ready-made paradigms for security professionals to understand these risks and introduce mitigation measures. The problem of a lack of recallability with autonomous weapons is similar to the problem of persistent mines. An approach along the lines of the U.S. landmine policy under the George W. Bush Administration, which required all land mines to self-destruct after a period of 4 hours to 15 days, and a dual-safe self-deactivating function after 90 days, would be a sensible policy response to bound this risk.\footnote{“U.S. Landmine Policy,” U.S. Department of State, February 27, 2004, \url{https://2001-2009.state.gov/t/pm/wra/c11735.htm}; U.S. State Department, “Landmine Policy White Paper,” February 27, 2004, \url{https://2001-2009.state.gov/t/pm/rls/fs/30047.htm}; and U.S. State Department, “Frequently Asked Questions on the New United States Landmine Policy,” February 27, 2004, \url{https://2001-2009.state.gov/t/pm/rls/fs/30050.htm}.} Little would be lost in terms of operational benefit by introducing such a precaution, since the military situation on the ground would have evolved and the autonomous weapon could return to base for further guidance from human controllers, in any case. Such a precaution would also likely add only marginal costs to autonomous weapons, which would already rely on complex software. If this were the only risk from deploying fully autonomous weapons, it seems likely the international community would be able to address this concern through international regulation, such as a rule that fully autonomous weapons must self-deactivate after a set period of time. Moreover, there would be only weak incentives to cheat such a regulatory framework, since there would be little military benefit gained in doing so and establishing time limitations would, in fact, enhance militaries’ own control over their autonomous weapons. A lack of recallability poses a real problem for war termination, but one that is manageable through a combination of smart weapon system design, policy measures to mitigate risk, and international regulation to coordinate national responses.

The efficacy of these mitigation measures depends, however, on autonomous weapons functioning as intended. War termination would be significantly complicated if policymakers had to additionally worry about autonomous weapons malfunctioning or
performing in ways other than intended while they were outside of human control and not recallable. If unpredictable actions were a normal feature of autonomous weapons, then these mitigation measures would be insufficient. An inability to recall malfunctioning autonomous weapons is a far more difficult problem.

The remaining five mechanisms by which fully autonomous weapons might affect stability center around the degree to which autonomous systems will behave in a manner consistent with human operators’ intentions in real-world military operational environments. If they behave as operators intend, then they may increase stability by giving policymakers more control over their military forces and reducing the risk of human subordinates taking unauthorized escalatory actions. On the other hand, the inability of autonomous weapons to deviate from their programming could be problematic in a number of situations. Their rigid adherence to programming could be a detriment if facing situations that were not anticipated by human operators. Autonomous weapons would lack the ability to apply judgment to a situation, reason about commander’s intent, and flexibly respond in ways that might deviate from their programming. This need not be due to a failure in the weapon system itself. Rather, it could also come from a failure of the human designer or operator to anticipate the situation the autonomous weapon faces. War is unpredictable, and the same is true for crises when nations stand on the brink of war. Many situations in crises and war require subordinates to flexibly adapt to novel situations, including enemy adaptation or unanticipated environmental conditions. Autonomous weapons’ inability to do so effectively could lead to brittle behavior that could be problematic in a conflict or crisis, with autonomous weapons undertaking actions and escalating tensions in ways that were not intended by human operators or national policymakers.

Autonomous weapons need not lead to accidental full-scale war to be problematic from the perspective of stability. Accidents with autonomous weapons that escalated tensions could be extremely problematic, forcing policymakers to take stronger actions due to domestic political concerns or for fear of backing down and looking weak internationally. Conversely, scholarship by Dan Reiter has shown that in many instances an important brake on conflict is policymakers’ desire not to be the one to land the first blow, for fear of international condemnation.324 An accidental with a lethal autonomous weapon could provide a casus belli for a leader who desired war, giving them a reason to retaliate forcefully.

At the height of the Cuban missile crisis, Soviet Premier Nikita Khrushchev sent an impassioned letter to President Kennedy calling on them to work together to step back from the brink of nuclear war:

Mr. President, we and you ought not now to pull on the ends of the rope in which you have tied the knot of war, because the more the two of us pull, the tighter that knot will be tied. And a moment may come when that knot will be tied so tight that even he who tied it will not have the strength to untie it, and then it will be necessary to cut that knot, and what that would mean is not for me to explain to you, because you yourself understand perfectly of what terrible forces our countries dispose.\(^{325}\)

Autonomous weapons would not take away all human decision-making in crises, but they do have the potential to tighten the knot, perhaps so far that it cannot be undone.

The crux of the issue is whether or not fully autonomous weapons will be likely to act consistent with human intent in realistic military operational environments. If not, then widespread autonomous weapons would undermine crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could escalate conflicts or make it more difficult to end wars (hypothesis H1). If autonomous weapons were likely to operate consistent with human intent in realistic military operational environments, however, then widespread autonomous weapons would improve crisis stability, escalation control, and war termination by increasing political leaders’ control over their military forces and reducing the risk of accidents or unauthorized escalatory actions (hypothesis H2).

The next section will examine the extent to which autonomous weapons are likely to function as intended in realistic military operational environments. Over 25 case studies are examined to understand the ability of organizations to manage risks from complex, highly-automated systems in uncontrolled and adversarial environments. While there are some examples of organizations successfully managing risk from complex, automated systems, including some military examples, these cases do not apply well to realistic wartime environments. The case studies as a whole suggest that accidents would be a normal and inescapable feature of autonomous weapons in crises.

\(^{325}\) Department of State Telegram Transmitting Letter From Chairman Khrushchev to President Kennedy, October 26, 1962, [http://microsites.jfklibrary.org/cmc/oct26/doc4.html](http://microsites.jfklibrary.org/cmc/oct26/doc4.html).
and conflict. While accidents would not necessarily be common, their occurrence would not be surprising. Smart design, testing, training, and other measures could reduce the risk of accidents, but not eliminate them entirely. Moreover, the nature of these accidents would be such that engineers and policymakers would be hard-pressed to accurately estimate the probability of accidents in advance. The accuracy of reliability estimates, such as a determination that a system is 99.99% reliable, drawn from testing environments should be greeted with caution in estimating real-world performance. This does not necessarily mean that autonomous weapons would malfunction or act in unintended ways in all instances nor that they would contribute to escalation in all crises. Nor would accidents with autonomous weapons necessarily led to full-scale war. But on the whole, the track record of complex, automated and autonomous systems in both military and non-military settings suggests that autonomous weapons would overall be detrimental to stability.
PART IV: CONTROLLING AUTONOMOUS WEAPONS: RISK, PREDICTABILITY, AND HAZARD

Are autonomous weapons likely to function as intended in realistic military operational environments? To answer this question, Part IV applies two competing theories about risk and hazard in complex systems: (1) normal accident theory, which predicts that accidents are a normal feature of complex, tightly-coupled systems; and (2) the theory of high-reliability organizations, which predicts that organizations can operate complex, tightly-coupled systems with extremely low accident rates in the right settings, if the right bureaucratic incentives prevail. Because there is no sufficiently large body of case studies of fully autonomous weapons, which generally do not yet exist (with some narrow exceptions, such as the Harpy), Part IV will explore over 25 related case studies of military and non-military complex systems in real-world environments to generate insights on the ability of militaries to successfully manage the risks of autonomous weapons. These case studies include: complex but non-autonomous military systems such as submarines and nuclear weapons; non-military autonomous systems such as self-driving cars and automated stock trading; and the U.S. military’s experience with two existing semi-autonomous and supervised autonomous weapons, the Army’s Patriot air defense system and the Navy’s Aegis weapon system. These case studies illuminate how risk, predictability, and hazard change as systems become more autonomous as well as the unique features of military operating environments and how they affect predictability. Part IV concludes that normal accident theory best describes the situation of fully autonomous weapons and that militaries are highly unlikely to be able to achieve high-reliability with fully autonomous weapons. Unanticipated lethal actions are likely to be normal consequences of using fully autonomous weapons and militaries can reduce but not entirely eliminate these risks.

Understanding Accident Risk: Normal Accident Theory vs. High-Reliability Organizations

Normal accident theory

John Borrie has written on autonomous weapons and risk, drawing on a body of work surrounding hazard and accidents known as normal accident theory.\(^326\) Borrie has

argued, “with very complex technological systems that are hazardous—and I think autonomous weapons fall into that category of hazard because of their intended lethality … we have difficulty being able to say that we can remove the risk of unintentional lethal effects.” Borrie has compared autonomous weapons to complex, hazardous systems in other industries. Humans have decades of experience designing, testing, and operating complex systems for high-risk applications, from nuclear power plants to commercial airliners to spacecraft. Because of these experiences, there is a robust field of research on how to improve safety and resiliency in these systems. The experience with complex systems to date suggests that 100% error-free operation is extremely rare, however. In sufficiently complex systems, it is impossible to test every possible system state and combination of states; some unanticipated interactions are inevitable. Failures may be unlikely, but over a long enough timeline they are inevitable. Engineers refer to these incidents as “normal accidents” because their occurrence is inevitable, even normal, in complex systems. “Why would autonomous systems be any different?” Borrie asked.

The archetype of a normal accident is the Three Mile Island nuclear power plant meltdown in 1979. The Three Mile Island incident was a “system failure,” meaning that the accident was caused by the interaction of many small, individually manageable failures interacting in an unexpected and dramatic way. Small problems combined in unexpected, nonlinear ways that weren’t predicted ahead of time because of the sheer complexity of the system.

Moreover, nuclear reactors are tightly coupled, as are many other complex machines. Tight coupling is when an interaction in one component of the system directly and immediately affects components elsewhere. There is very little “slack” in the system – little time or flexibility for humans to intervene and exercise judgment, bend or break rules, or alter the system’s behavior. It is the combination of system complexity and tight coupling that makes accidents normal. In loosely coupled complex systems, such as bureaucracies or other human organizations, there is sufficient slack for humans to adjust to unexpected situations and manage failures. In tightly coupled

(Note: The author participated in a UNIDIR-hosted workshop that helped inform Borrie’s project and spoke at a UNIDIR-hosted panel in 2016.)

327 John Borrie, Research Coordinator and Programme Lead, UN Institute for Disarmament Research, interview by author, Geneva, Switzerland, April 12, 2016.

328 Ibid.
systems, however, failures can rapidly cascade from one sub-system to the next and minor problems can quickly lead to system breakdown.

Borrie argued that autonomous weapons would have the same characteristics of complexity and tight coupling, making them susceptible to “failures … we hadn’t anticipated.”329 A brief overview of the Three Mile Island meltdown can help shed light on how failures occur in complex systems.330

The Three Mile Island meltdown

The problems at Three Mile Island began in the early hours of March 28, 1979, with a minor and seemingly insignificant equipment failure. A seal began to leak in a water cooling system. This, in itself, did nothing. The water cooling system remained functional and, in any case, was not critical to the reactor’s operation. But the leaky seal set in motion a chain of events that would result in the worst nuclear disaster in American history.

Moisture from the leaky seal got into an unrelated system, causing it to shut off two water pumps vital to cooling the reactor. This was an unexpected interaction between two otherwise unrelated components of the reactor. On paper, the leaky seal should not have led to the water pumps shutting down. In practice, however, the two seemingly unrelated sub-systems were physically located near each other, even if unconnected.

The water pumps shutting down was a manageable problem. Automated safeties kicked in, turning off the turbines that generate electric power and activating emergency pumps. Nuclear reactors are hot and require a constant flow of water to help shed their heat in a safe manner. So long as water was flowing, the reactor would remain safe.

However, a valve needed to allow water to flow through the emergency cooling system had been left closed. This was a case of operator error. So a minor equipment failure plus an unexpected interaction between parts of the system and human error combined to create a situation where vital water wasn’t flowing to cool the reactor.

329 Ibid.

330 This description is taken from Perrow, Normal Accidents: Living with High-Risk Technologies, 15–31; and United States Nuclear Regulatory Commission, “Backgrounder on the Three Mile Island Accident.”
Human operators were managing the reactor, observing it in real time and could have responded to address this problem. But, through no fault of their own, they were unaware that the valve was shut. The indicator light on their control panel showing the valve’s position was obscured by a repair tag for another, unrelated system.

Now one begins to see the essence of normal accidents – the combination of individually manageable issues: minor equipment failure plus unexpected interactions plus human error plus poor system design. Not only was vital water not flowing to cool the reactor, the human operators weren’t even aware of the problem.

Without water, the reactor core temperature began to rise, which activated another automatic safety. The reactor “scrammed,” dropping graphite control rods into the reactor core to absorb neutrons and stop the chain reaction. This stopped the nuclear reaction. However, the core was still generating heat. Constant water flow was required to help safely shed heat, but both the primary and secondary water pumping systems had failed.

As the core temperature rose, the pressure inside the containment vessel began to rise as well. Excessive pressure is very dangerous as it can cause the containment vessel to crack, releasing dangerous radiation into the air. In response to rising pressure, another automatic safety kicked in: an automated pressure-release valve. This was designed to let off steam if the pressure got too high.

The automated pressure-release valve opened as designed, letting off steam. Unfortunately, the valve failed to close. Moreover, the valve’s indicator light also failed, so the plant’s operators did not know the valve was stuck open. On top of the failures that had already occurred, these additional ones proved disastrous. With the valve stuck open, too much steam was released and the water levels in the reactor core began to fall. Almost a third of the water in the reactor escaped. Because water was crucial to cooling the still-hot nuclear core, another automatic emergency water cooling system kicked in. The plant’s operators also activated an additional emergency water cooling system.

This sequence of events occurred very quickly. Everything described thus far occurred within a mere 13 seconds. This highlights the tight coupling of complex systems. Seemingly minor issues, such as a pressure release valve stuck open, led directly and swiftly to very dangerous conditions.

In this short span of time, multiple automatic safeties kicked in which helped to mitigate the danger. Additionally, the human operators in charge of Three Mile Island were not passive observers. They reacted quickly in response to what they were seeing.
at their control panels. But in their responses we see the limitations of both the humans and automatic safeties.

The automatic safeties were useful, but did not fully address the root causes of the problems – an emergency water cooling valve that was closed when it should have been open and a pressure-release valve that was stuck open when it should have been closed. In principle, “smarter” safeties that took into account more variables could have addressed these issues. Indeed, nuclear reactor safety has improved considerably since Three Mile Island.

The human operators faced a different problem, however, one which more sophisticated automation actually makes harder, not easier: the incomprehensibility of the system itself. Because the human operators could not directly inspect the internal functioning of the reactor core, they had to rely on indicators to tell them what was occurring. But these indicators were also susceptible to failure, and some of them did fail. In the case of the two chief problems – the water cooling valve that was closed and the pressure-release valve that was stuck open – one indicator was obscured by an ill-placed repair tag and the other was malfunctioning. In their attempts to manage the situation, the human operators suffered from a substantial deficit of information about the system and its internal workings. The operators did not discover that the emergency water cooling valve was improperly closed until eight minutes into the accident and that the pressure release valve was stuck open until two hours and twenty minutes into the accident. This meant that some of the corrective actions they took to manage the accident were, in retrospect, incorrect. It would be improper to call their actions “human error,” however. They were operating with the best information they had at the time. They simply could not have known better.

The father of normal accident theory, Charles Perrow, points out that the “incomprehensibility” of complex systems themselves is a stumbling block to predicting and managing normal accidents. The system is so complex that it is incomprehensible, or opaque, to users and even the system’s designers. This problem is exacerbated in situations in which humans cannot directly inspect the system, such as a nuclear reactor, but also exists in situations where humans are physically present. During the Apollo 13 disaster, it took 17 minutes for the astronauts and NASA ground control to uncover the source of the instrument anomalies they were seeing, in spite of the fact that the astronauts were on board the craft and could “feel” how the spacecraft was performing. The astronauts heard a bang and felt a small jolt from the initial explosion in the oxygen tank and could tell that they had trouble controlling the attitude (orientation) of the craft. Nevertheless, the system was incomprehensible enough that
vital time was lost as the astronauts and ground control experts poured over the various instrument readings and rapidly-cascading electrical failures before they discovered the root cause.331

The inevitability of accidents

The Apollo 13 and Three Mile Island incidents date from the 1970s, when engineers were still learning to manage complex, tightly-coupled systems. Since then, both nuclear power and space travel have become safer and more reliable. They can never be made entirely safe, however.

NASA has seen additional tragic accidents, including some that were not recoverable as Apollo 13 was. These include the loss of the space shuttles Challenger (1986) and Columbia (2003) and their crews. While these accidents had discrete causes that could be addressed in later designs (faulty O-rings and falling foam insulation, respectively), the impossibility of anticipating these specific failures in advance makes continued accidents inevitable, if rare.332 In 2015, for example, the private company SpaceX had a rocket blow up on the launch pad due to a strut failure that had not been previously identified as a risk. A year later, another SpaceX rocket exploded during testing due to a problem with supercooled oxygen that SpaceX CEO Elon Musk said had “never been encountered before in the history of rocketry.”333

Nuclear power has similarly increased significantly in safety, but the 2011 meltdown of the Japanese Fukushima Daiichi nuclear plant points to the limits of safety. Fukushima Daiichi was hardened against earthquakes and flooding, with backup generators and 30-foot high flood walls. Unfortunately, the plant was not prepared for a 9.0 magnitude earthquake (the largest recorded earthquake to ever hit Japan) off the

---

331 For a very brief summary of the incident, see National Aeronautics and Space Administration, “Apollo 13.” NASA’s full report on the Apollo 13 disaster can be found at National Aeronautics and Space Administration, “Report of the Apollo 13 Review Board.” See also Perrow, Normal Accidents, 271–281.

332 On Challenger, see National Aeronautics and Space Administration, “Report of the Presidential Commission on the Space Shuttle Challenger Accident.” On the Columbia accident, see National Aeronautics and Space Administration, “Columbia Accident Investigation Board, Volume 1.”

coast. This quake caused both a loss in power and a massive 40-foot high tsunami. Many safeties worked correctly. The earthquake did not damage the containment vessels. When the earthquake knocked out primary power, the reactors automatically scrammed, inserting control rods to stop the nuclear reaction. Backup diesel generators automatically came online.\textsuperscript{334}

However, the 40-foot high tsunami wave topped the 30-foot high flood walls, swamping 12 of 13 backup diesel generators. Combined with the loss of primary power from the electrical grid, the plant lost the ability to pump water to cool the still-hot reactor cores. Despite the heroic efforts of Japanese engineers to bring in additional generators and pump water into the overheating reactors, the result was the worst nuclear power accident since Chernobyl.\textsuperscript{335}

The problem wasn’t that Fukushima Daiichi lacked backup safeties. The problem was a failure to anticipate an unusual environmental condition (a massive earthquake off the coast that induced a tsunami) that caused a common-mode failure—one that simultaneously overwhelmed two seemingly independent safeties: primary and backup power. Even in fields where safety is a central concern, such as space travel and nuclear power, anticipating all of the possible interactions of the system and its environment is effectively impossible.

The problem with complex systems isn’t just that failures are inevitable. It’s that the sheer complexity of systems inhibits predicting when and how failures are likely to occur. If failures can be predicted, they can be accounted for. In simple systems, all of the components can be rigorously tested so that the system’s limits can be understood. When the limits are understood, they can be taken into account, with humans modifying how they use the system. In sufficiently complex systems, however, unexpected interactions within the system, between the system and its environment, and between the system and human operators are bound to occur. These challenges come up time and again in complex, tightly-coupled systems and autonomous weapons will have the same vulnerabilities.

\textsuperscript{334} Lipscy et al, “The Fukushima Disaster and Japan’s Nuclear Plant Vulnerability in Comparative Perspective.”

\textsuperscript{335} Ibid.
High-reliability theory

Is safe operation of hazardous complex systems impossible? Normal accident theory would suggest that these systems can be made safer, but never 100% safe. The probability of accidents can be reduced, but never eliminated.

There is an alternate point of view on complex systems, however, which suggests that under certain conditions, normal accidents can largely be avoided. “High-reliability organizations” are those that exhibit certain characteristics that allow them to routinely operate high-risk systems with very low accident rates. Examples of high-reliability organizations include the FAA air traffic control system and U.S. Navy aircraft carrier flight decks. These organizations do not have zero accident rates, but they are exceptionally low given the complexities of their operating environment and the hazards of operation. High-reliability organizations can be found across a range of industries but exhibit certain common characteristics. These include not only highly-trained individuals, but also a collective mindfulness of the risk of failure and a continued commitment to learn from near misses and improve safety.

While militaries as a whole would not be considered high-reliability organizations, some select military communities have very high safety records with complex high-risk systems. In addition to aircraft carrier flight deck operations, the U.S. Navy’s submarine community is an example of a high-reliability organization. Following the loss of the USS Thresher to an accident in 1963—at the time one of the Navy’s most advanced submarines and first in her class—the Navy instituted the SUBSAFE program to improve submarine safety. SUBSAFE is a continuous process applied to design, material, fabrication, and testing to ensure safe submarine operations. In Congressional testimony in 2003, Rear Admiral Paul Sullivan, the Navy deputy commander for ship design, integration, and engineering, explained the impact of the program:


337 Weick and Sutcliffe, Managing the Unexpected: Sustained Performance in a Complex World.

The SUBSAFE Program has been very successful. Between 1915 and 1963, 16 submarines were lost due to non-combat causes, an average of one every three years. Since the inception of the SUBSAFE Program in 1963, only one submarine has been lost. USS _Scorpion_ (SSN 589) was lost in May 1968 with 99 officers and men aboard. She was not a SUBSAFE certified submarine and the evidence indicates that she was lost for reasons that would not have been mitigated by the SUBSAFE Program. We have never lost a SUBSAFE certified submarine.\(^{339}\)

SUBSAFE is a shining example of what high-reliability organizations can do. Operating a nuclear-powered submarine is exceptionally complex and inherently hazardous, and yet the Navy has been able to substantially reduce these risks. Accidents resulting in catastrophic loss of a submarine are not “normal” in the U.S. Navy. Indeed, they are unprecedented since the advent of SUBSAFE. It would appear that, at least in some circumstances, high-reliability organizations can achieve a level of safety that refutes the main premise of normal accident theory.

 Applying normal accident theory and high-reliability to autonomous weapons

Normal accident theory explains the persistence of accidents even in industries that are heavily invested in safety, such as nuclear power plants and space travel. Normal accident theory predicts that safety precautions can reduce the probability of accidents, but never entirely eliminate them. At the same time, there are some industries that appear to defy the predictions of normal accident theory. These include the FAA air traffic control system, U.S. aircraft carrier deck operations, and U.S. Navy submarine operations. In some cases, the safety record in these industries is astounding. U.S. air carriers have had one passenger fatality in nine years, despite flying 100 million flights per year carrying billions of passengers.\(^{340}\) The U.S. Navy has never lost a SUBSAFE-certified submarine.

The difference between high-reliability and normal accidents is principally in the characteristics of the organization operating a complex, hazardous system. Air travel

---


and submarine operations are inherently hazardous. U.S. air carriers and the U.S. Navy have achieved high reliability operations by instituting strict organizational procedures that successfully mitigate these risks. When other organizations have not, their accident rates are more consistent with what normal accident theory predicts. The Russian/Soviet Navy, for example, has a fair higher submarine accident rate.\textsuperscript{341}

Autonomous weapons have hazardous characteristics. They are complex, tightly-coupled, lethal systems. The question is whether these risks can be successfully managed to achieve high-reliability in spite of their inherent hazard. Autonomous weapon systems are different from the complex mechanical systems described above (nuclear power plants, spacecraft, submarines) in a number of respects. Much of autonomous weapons’ complexity is due to software; they have autonomous features; and they operate in competitive, uncontrolled environments. To better understand how these factors would affect reliability, we will examine over a dozen case studies of systems with similar characteristics. These include non-military autonomous systems in competitive and uncontrolled environments and military automated systems.

Case Studies of Non-Military Autonomous Systems in Competitive and Uncontrolled Environments

\textit{Safety in complex software-based systems}

Many modern complex systems use a significant amount of software and automation. Many of these features significantly improve safety, and in fact in many cases automation is responsible for improvements in safety over the past several decades, such as in commercial airliners. Modern jetliners are effectively capable of flying themselves, with pilots functioning largely as an emergency backup option. Modern automobiles still have human drivers, but have incorporated a host of automated or autonomous features to improve driving safety and comfort: anti-lock brakes, traction and stability control, automatic lane keeping, intelligent cruise control, collision avoidance, and self-parking. Even modern fighter jets use software to help improve safety and reliability. F-16 fighter aircraft have been upgraded with automatic ground collision avoidance systems (Auto-GCAS) to prevent crashes in the event that pilots lose situational awareness. Auto-GCAS has already saved at least one life when it

\textsuperscript{341} For a list of catastrophic Soviet submarine accidents, see “Lost submarines of the Soviet Union,” Wikipedia, 

165
prevented an F-16 from crashing in Syria. The more advanced F-35 fighter reportedly has software-based limits on its flight controls to prevent pilots from putting the aircraft into unrecoverable spins or other aerodynamically unstable conditions. These automated features improve safety.

The double-edged sword to this automation is that all of this added software increases complexity, which can itself introduce new problems. Sophisticated automation requires software with millions of lines of code: 8 million for the F-22 fighter jet, 24 million for the F-35 jet, and some 100 million lines of code for a modern luxury automobile. All things being equal, longer pieces of software are harder to verify that they are free from bugs or glitches. Studies have pegged the software industry average error rate at 15 to 50 errors per 1000 lines of code. Rigorous test and evaluation has reduced the error rate to 0.1 to 0.5 errors per 1000 lines of code in some cases. However, in systems with millions of lines of code, some errors are inevitable. If they aren’t caught in testing, the results can be unfortunate if they are discovered during real world operation.

On their first deployment to the Pacific in 2007, eight F-22 fighter jets experienced a Y2K-like total computer meltdown when crossing the international dateline. All onboard computer systems crashed, causing the pilots to lose navigation, fuel subsystems, and some communications. Stranded over the Pacific without a navigational reference point, the aircraft were able to make it back to land by following the tanker aircraft accompanying them, which relied on an older computer system. Under tougher circumstances, such as combat or even bad weather, the incident could

---


346 Sheppard, “This Week at War.”
have led to a catastrophic loss of the aircraft. While the existence of the international
dateline clearly could be anticipated, the interaction of the dateline with the software
was not identified in testing.

The October 2018 and March 2019 crashes of the Boeing 737 MAX aircraft,
which killed a total of 346 people, are another example of a software-related accident
early in a complex system’s operational fielding. While the commercial airline industry
is highly regulated and planes are rigorously tested, the 737 MAX incidents
nevertheless stand out as stark examples of normal accidents with automated systems.
Many bureaucratic failures contributed to the fielding of a flawed automated flight
control system, the Maneuvering Characteristics Augmentation System (MCAS), which
instigated both crashes. The MCAS was originally designed for one purpose, then
repurposed to a more consequential role, but not fully redesigned. Backup sensors were
eliminated but without a reexamination of the safety implications. Some assessments of
risk were massively wrong. An erroneous activation of MCAS was estimated to occur
in less than one in 10 million flight hours, but that didn’t take into account external
events like bird strikes or mechanics bumping the sensors, which happen routinely in
the real world. This meant the real-world odds of an accident were much, much higher
than testers estimated.\(^{347}\) Despite a company with world-class engineers, a brand to
protect, and an outside regulatory agency with oversight responsibilities, the company
still fielding an aircraft that had fatal flaws in real world environments.

The 737 MAX incidents illustrate the limits of safety even in industries that
have extremely high reliability. High reliability is achieved through routine operations,
so it is no surprise that the 737 MAX crashes occurred with a relatively new automated
flight control software that had been in operation for less than 18 months in real-world
environments. While high-reliability operations may eventually be achieved with a
revised version of the 737 MAX and additional operational experience, high-reliability
should not be expected during the early phases of real-world use from a complex
automated system in a high-risk setting. Under non-routine conditions, normal accident
theory applies.

Software vulnerabilities can bring other challenges, such as creating
opportunities for hackers. In 2015, two hackers revealed vulnerabilities that allowed
them to remotely hack certain automobiles while they were on the road. This allowed

\(^{347}\) Jack Nicas et al., “Boeing Built Deadly Assumptions Into 737 MAX, Blind to a Late Design
them to take control of critical driving functions including the transmission, steering, and brakes.348

Even if software does not have specific bugs or vulnerabilities, the sheer complexity of modern machines can make it challenging for users to understand what the automation is doing and why. When humans are no longer interacting with simple mechanical systems that may behave in predictable ways but instead are interacting with complex pieces of software with millions of lines of code, the human user’s mental model for how the automation is behaving may diverge significantly from what it is actually doing. While users might be able to build an accurate mental model for how a simple system works, such as an old mechanical thermostat, the internal functioning and reasoning of more complex systems such as a self-driving car may be more opaque to the user. Borrie observed, “As systems get increasingly complex and increasingly self-directed, I think it's going to get more and more difficult for human beings to be able to think ahead of time what those weak points are necessarily going to be.”349 This opacity of complex systems can also contribute to accidents.

“We don’t understand anything!” Human-machine interaction failures

An example of an accident due human-machine interaction failures is the Air France Flight 447 crash in June 2009. Four hours into the flight, the aircraft suffered a minor and insignificant instrumentation failure when air speed probes on the wings froze over due to ice crystals, a rare but non-serious problem. (It was a known problem and the manufacturer, Airbus, was in the process of replacing the air speed probes.)

The clogging of the air speed probes with ice crystals did not affect the flight of the aircraft, but it did cause airspeed indicators in the cockpit to register improbably low levels. Faced with this instrumentation failure, the autopilot disengaged and handed over control back to the pilots. Significantly, the plane also entered a different software mode for flight controls. Instead of flying under “normal law” mode, where software limitations prevent pilots from putting the plane into dangerous aerodynamic conditions such as stalls, the plane entered “alternate law” mode, where the software limitations are relaxed and the pilots have more direct control over the plane.

Nevertheless, there was no actual emergency. Eleven seconds following the autopilot disengagement, the pilots correctly identified that they had lost the airspeed


349 Borrie, interview.
indicators. Because this was only an instrumentation problem, it did not directly affect the flight of the aircraft. At this point, the aircraft was flying normally, at appropriate speeds and full altitude, and there was no problem. Aerodynamically, everything was fine.

Inexplicably, however, the pilots began a series of errors that resulted in a stall, causing the aircraft to crash into the ocean, killing all 228 people on board. Throughout the incident, the pilots continually misinterpreted data from the airplane and misunderstood the aircraft’s behavior. At one point mid-crisis, the co-pilot exclaimed, “We completely lost control of the airplane and we don’t understand anything! We tried everything!” The problem was actually simple. The pilots had pulled back on the stick, causing the aircraft nose to pitch too high. This resulted in a stall where the wings lose lift, causing the airplane to rapidly lose altitude. This is a basic aerodynamic concept and the solution is to bring the nose down, putting the plane into a dive to bring lift back into the wings. Instead, the pilots kept pulling back even further on the stick, bringing the nose higher and higher, keeping the plane in a stall. In part, poor user interfaces and opaque automated processes on the aircraft, even while flown manually, contributed to the pilots’ lack of understanding. The complexity of the aircraft created problems of transparency that would likely not have existed in a similar situation on a simpler aircraft. By the time the senior pilot understood what was happening, it was too late. The plane was too low and descending too rapidly. It was past the point of recovery. The plane crashed into the ocean, killing everyone onboard.350

Unlike in the F-22 international dateline incident, 737 MAX crashes, or the automobile hack, the Air France Flight 447 crash was not due to a hidden vulnerability lurking within the software. The automation performed perfectly. It would be overly simplistic to lay the crash at the feet of human error, however. The pilots made mistakes, but the problem is best characterized as human-automation failure. The pilots were confused by the automation and the complexity of the system.

Bot vs. Bot

Another source of failure in automated systems in uncontrolled environments can be interactions with other automated systems. An example of this phenomenon is when two warring bots elevated the price of an otherwise ordinary book on Amazon to $23 million.\footnote{Michael Eisen, “Amazon’s $23,698,655.93 book about flies,” it is NOT junk, April 22, 2011, http://www.michaieleisen.org/blog/?p=358.}

Two online sellers, bordeebook and profnath, both of whom were legitimate online booksellers with thousands of positive ratings, became locked in a runaway price war for an otherwise unremarkable biology textbook, Peter Lawrence’s *Making of a Fly: The Genetics of Animal Design*. Once a day, profnath would set its price to 0.9983 times bordeebook’s price, slightly undercutting them. A few hours later, bordeebook would change its price, increasing it to 1.270589 times profnath’s price. The combined effect was to increase bordeebook’s price by approximately 27% each day. The following day, profnath would slightly undercut bordeebook’s new price and the prices would escalate.\footnote{Ibid.}

The price peaked out at $23,698,655.93 (plus $3.99 shipping) before dropping back to a more tame $134.97, where it stayed. (Perhaps because someone at one of the companies noticed and intervened.) In a blog post on the incident, Michael Eisen mused about the possibilities for “chaos and mischief” that this discovery suggested.\footnote{Ibid.} A person could hack this vulnerability of the bots, manipulating prices. In fact, unanticipated interactions between bots and deliberate manipulation have both been seen in the far more consequential domain of stock trading.

*Automated stock trading and “flash crashes”*

trading has similar characteristics to fully autonomous weapons: highly complex software operating at machine speed in an unpredictable environment populated by adversarial automated systems. Automated stock trading is therefore an ideal case study for examining risk and hazard under these conditions. Multiple accidents with trading algorithms suggest that militaries are likely to experience accidents with fully autonomous weapons in real-world settings.

Automated stock trading, sometimes called algorithmic trading, is when computer algorithms are used to monitor the market and make trades based on certain conditions.\(^{356}\) The simplest kind of algorithm, or “algo,” is used to break up large trades into smaller ones in order to minimize the costs of the trade. If a single buy or sell order is too large relative to the volume of that stock that is regularly traded, placing the order all at once can skew the market price. To avoid this, traders use algorithms to break up the sale into pieces that can be executed incrementally according to stock price, time, volume, or other factors. In such cases, the decision to make the trade (to buy or sell a certain amount of stock) is still made by a person. The machine simply handles the execution of the trade.

Some trading algorithms take on more responsibility, actually making automated trading decisions to buy or sell based on the market.\(^{357}\) For example, an algorithm could be tasked to monitor a stock’s price over a period of time. When the price moves significantly above or below the average of where the price has been, the algo sells or buys accordingly, under the assumption that over time the price will revert back to the average, yielding a profit. Another strategy could be to look for arbitrage opportunities, where the price of a stock in one market is different from the price in another market, and this price difference can be exploited for profit. All of these strategies could, in principle, be done by humans. Automated trading offers the

---


357 Some writers on automated stock trading differentiate between automated trading and algorithmic trading, using the term algorithmic trading only to refer to the practice of breaking up large orders to execute via algorithm by price, time, or volume, and referring to other practices such as seeking arbitrage opportunities as automated trading. Others treat algorithmic trading and automated trading as effectively synonymous.
advantage, however, of monitoring large amounts of data and immediately and precisely making trades in ways that would be impossible for humans.\textsuperscript{358}

Speed is a vital factor in stock trading. If there is a price imbalance and a stock is under- or over-priced, there are many other traders also looking to sweep up that profit. Move too slow and one could miss the opportunity. The result has been an arms race in speed and the rise of high-frequency trading, a specialized type of automated trading that occurs at speeds too quick for humans to even register. High-frequency trades move at speeds measured in microseconds: \text{0.000001 seconds.}\textsuperscript{359} The blink of an eye takes a fraction of a second – 0.1 to 0.4 seconds – but is still an eon compared to high-frequency trading. During the span of a single eye blink, 100,000 microseconds pass by. The result is an entirely new ecosystem, a world of trading bots dueling at superhuman speeds only accessible by machines.

The gains from even a slight advantage in speed are so significant that high-frequency traders will go to great lengths to shave just a few microseconds off their trading times. High-frequency traders co-locate their servers within the server rooms of stock exchanges, cutting down on travel time. Some are even willing to pay additional money to move their firm’s servers a few feet closer to the stock exchange’s servers within the room. Firms try to find the shortest route for their cables within the server room, cutting microseconds off of transit time.\textsuperscript{360} High-frequency traders spare no expense in optimizing every part of their hardware for speed, from data switches to the glass inside fiber optic cables.\textsuperscript{361}

At the time scales at which high-frequency trading operates, humans have to delegate trading decisions to the algorithms. Humans can’t possibly observe the market and react to it in microseconds. That means if things go wrong, they can go wrong very quickly. To ensure algorithms do what they are designed to do once released into the


\textsuperscript{360} Ibid, 62-63.

\textsuperscript{361} Ibid, 63-64.
real world, developers test them against actual stock market data, but with trading disabled.\textsuperscript{362} Despite this, accidents still occur.

\textit{“Knightmare on Wall Street”}

In 2012, Knight Capital Group was a titan of high-frequency trading. Knight was a “market maker,” a high-frequency trader that executed over 3.3 billion trades, totaling $21 billion, every single day. Like most high-frequency traders, Knight didn’t hold on to this stock. Stocks were bought and sold the same day, sometimes within fractions of a second. Nevertheless, Knight was a key player in the U.S. stock market, executing 17\% of all trades on the New York Stock Exchange and NASDAQ. Their slogan was, “The Science of Trading, the Standard of Trust.”\textsuperscript{363} Like many high-frequency trading firms, their business was lucrative. On the morning of July 31, 2012, Knight had $365 million in assets. Within 45 minutes, it would all be gone.

At 9:30am Eastern Time on July 31st, U.S. markets opened and Knight deployed a new automated trading system. Instantly, it was apparent that something was wrong. One of the functions of the automated trading system was to break up large orders into smaller ones, which then would be executed individually. Knight’s trading system wasn’t registering that these smaller trades were actually completed, however, so it kept tasking them again, creating an endless loop of trades. Knight’s trading system began flooding the market with orders, executing over a thousand trades a second. Even worse, Knight’s algorithm was buying high and selling low, losing money on every trade.\textsuperscript{364}

There was no way to stop it. The developers had neglected to install a “kill switch” to turn their algorithm off.\textsuperscript{365} While Knight’s computer engineers worked to diagnose the problem, the software was actively trading in the market, moving $2.6


\textsuperscript{365} D7, “Knightmare.”
million a second. By the time they finally halted the system 45 minutes later, the runaway algo had executed 4 million trades, moving $7 billion. Some of those trades made money, but Knight lost a net $460 million. The company only had $365 million in assets. Knight was effectively bankrupt.

An influx of cash from investors helped Knight cover their losses, but the company was ultimately sold. The incident became known as the “Knightmare on Wall Street.” Knight’s runaway algo was frightening, but it was the risk of using an autonomous system in a high-stakes application. Accidents happen, and when they happen with autonomous systems with no ability for humans to intervene, they can cause significant damage. Despite their experience in high-frequency trading, Knight was taking risks with their automated stock trading system that they didn’t fully understand.

2010 Flash Crash

The damage from Knight’s trading debacle was largely contained to a single company, but one of the most notable automated trading accidents was the May 6, 2010 “flash crash.” A volatile combination of factors meant that during the flash crash, one malfunctioning algorithm interacted with an entire marketplace ready to run out of control. For a few brief minutes that afternoon, the U.S. stock market took a precipitous plunge, losing 10 percent of its value before recovering. The 2010 flash crash demonstrated the risks of automated trading, shook investor confidence, and led to new safety mechanisms by regulators to halt further incidents from spiraling out of control.


367 Ibid.

368 D7, “Knightmare.”


371 Waddell & Reed was not named in the official SEC and CFTC report, which referred only to a “large fundamental trader (a mutual fund complex).” U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission, Findings Regarding the Market
smaller type of futures contract, one-fifth the size of a regular futures contract. A futures contract is what it sounds like: an agreement to buy or sell at a certain price at a certain point in time in the future.) Because executing such a large trade all at once could distort the market, Waddell & Reed used a “Sell Algorithm” to break up the sale into smaller trades, a standard practice. The algorithm was tied to the overall volume of E-minis sold on the market, with direction to execute the sale at 9% of the trading volume over the previous minute. In theory, this should have spread out the sale so as to not overly influence the market.

The Sell Algorithm was given no instructions with regard to time or price, however, an oversight that led to a catastrophic case of brittleness. The market that day was already under stress. Government investigators later characterized the market as “unusually turbulent,” in part due to an unfolding European debt crisis that was causing uncertainty. By mid-afternoon, the market was experiencing “unusually high volatility” (sharp movements in prices) and low liquidity (low market depth). It was into these choppy waters that the Sell Algorithm sailed.

Only twice in the previous year had a single trader attempted to unload so many E-Minis on the market in a single day. Normally, a trade of this scale took hours to

---


376 Ibid, 2.
execute. This time, because the Sell Algorithm was only tied to volume and not price or time, it happened very quickly: within 20 minutes.

The Sell Algorithm provided the spark, and high-frequency traders were the gasoline. High-frequency traders bought the E-minis the Sell Algorithm was unloading and, as is their frequent practice, rapidly re-sold them. This increased the volume of E-minis being traded on the market. Since the rate at which the sell Algorithm sold E-minis was tied to volume but not price or time, it accelerated its sales, dumping more E-minis on an already stressed market.

Without buyers interested in buying up all of the E-minis that the Sell Algorithm and high-frequency traders were selling, the price of E-minis dropped, falling 3% in just four minutes. This generated a “hot potato” effect among high-frequency traders as they tried to unload the falling E-minis onto other high-frequency traders.\(^{377}\) In one 14 second period, high-frequency trading algorithms exchanged 27,000 E-mini contracts.\(^ {378}\) (For reference, the total amount Waddell & Reed were trying to sell was 75,000 contracts.) All the while, as trading volume skyrocketed, the Sell Algorithm kept unloading more and more E-minis on a market that was unable to handle them.

The plummeting E-minis dragged down other U.S. markets. Observers incredulously watched the Dow Jones, NASDAQ, and S&P 500 all inexplicably plunge. Finally, at 2:45:28 pm, an automated Stop Logic safety on the Chicago Mercantile Exchange kicked in, halting E-mini trading for 5 seconds and allowing the markets to reset.\(^ {379}\) They rapidly recovered, but the sharp distortions in the market wreaked havoc on trading. Over 20,000 trades had been executed at what financial regulators termed “irrational prices” far from their norm, some as low as a penny or as high as $100,000.\(^ {380}\) After the markets closed, the Financial Industry Regulatory Authority worked with stock exchanges to cancel tens of thousands of “clearly erroneous” trades.\(^ {381}\)

However, the repercussions from the “flash crash” were only beginning. Asian markets tumbled when they opened the next day, and while the markets soon stabilized, it was harder to repair confidence. Traders described the flash crash as “horrifying” and

---

\(^{377}\) Ibid, 3.

\(^{378}\) Ibid, 3.

\(^{379}\) Ibid, 4.

\(^{380}\) Ibid, 5.

\(^{381}\) Ibid, 6.
“absolute chaos,” reminiscent of the 1987 “Black Monday” crash where the Dow Jones lost 22% of its value. Market corrections had occurred before, but the sudden downward plunge followed by an equally rapid reset was something entirely different, an inhuman event driven by faulty machines interacting in ways that no one in the marketplace fully understood.

The flash crash demonstrated how brittle algorithms interacting with a complex environment can have unforeseen reactions. When algorithms interact with one another at superhuman speeds, the result can be a runaway process with catastrophic consequences. The stock market as a whole is an incredibly complex system that defies simple understanding, which can make predicting interactions difficult ahead of time. On a different day, under different market conditions, the same Sell Algorithm may not have led to a flash crash. In the investigations that have followed the flash crash, researchers have blamed everything from human error to brittle algorithms, high-frequency trading, market volatility, and deliberate market manipulation. In truth, all of them likely played a role. Like other normal accidents, the flash crash had multiple causes, any one of which individually would have been manageable. The combination, however, was uncontrollable.

Spoofing the bots

In addition to the risk of unanticipated interactions among algorithms, there is a further complication to the 2010 flash crash that illustrates another important risk: deliberate manipulation of bot behavior by humans. In April 2015, the U.S. Justice Department charged London-based trader Navinder Singh Sarao with fraud and market manipulation contributing to the 2010 flash crash. Sarao, they alleged, had manipulated the market by “spoofing” – sending false data. According to the Justice Department, Sarao used automated trading algorithms to place multiple large-volume orders to create the appearance of demand to drive up price, then cancelled the orders before they were

---


executed. By deliberately manipulating the price, Sarao could buy low and sell high, making a profit as the price moved.384

In November 2016, Sarao pled guilty to fraud and spoofing, admitting that he used an automated trading algorithm to manipulate the market for E-minis.385 The argument that he was the cause of the flash crash has been criticized by outside experts, however.386 For one, Sarao continued his alleged market manipulation for five years after the flash crash until finally arrested in 2015. The authorities didn’t accuse him of causing any other flash crashes. Perhaps more significantly, Sarao’s spoofing algorithm was reportedly turned off during the sharpest downturn in the flash crash.387 Sarao’s spoofing might have spooked the markets, contributing to instability and helping to start the downslide, but it would seem a stretch to conclude that he was the cause. In fact, the Justice Department’s specific wording alleges that Sarao only “contributed” to the flash crash.388 Nevertheless, Sarao’s involvement points to the risk of brittle algorithms being manipulated by competitors in adversarial environments.

“Circuit breakers” as safety measures

In the aftermath of the flash crash, regulators installed “circuit breakers” to limit future damage.389 Circuit breakers, which were first introduced after the 1987 Black

---


388 Department of Justice, “Futures Trader Charged with Illegally Manipulating Stock Market, Contributing to the May 2010 Market ‘Flash Crash.’”

389 The first tranche of individual stock circuit breakers, implemented in the immediate aftermath of the Flash Crash, initiated a five-minute pause if a stock’s price moved up or down more than 10 percent in the preceding five minutes. U.S. Commodity Futures Trading Commission and
Monday crash, halt trading if stock prices drop too quickly. Market-wide circuit breakers trip if the S&P 500 drops more than 7%, 13% or 20% from the closing price the previous day, temporarily pausing trading or, in the event of a 20% drop, shutting down markets for the day. After the flash crash, in 2012 the Securities and Exchange Commission introduced new “limit up – limit down” circuit breakers for individual stocks to prevent sharp, dramatic price swings. The limit up—limit down mechanism creates a price band around a stock, based on the stock’s average price over the preceding five minutes. If the stock price moves out of that band for more than 15 seconds, trading is halted on that stock for five minutes.

Circuit breakers are an important mechanism for preventing flash crashes from causing too much damage. We know this because they keep getting tripped. An average day sees a handful of circuit breakers tripped due to rapid price moves. One day in August 2015, over 1,200 circuit breakers were tripped across multiple exchanges. Mini-flash crashes have continued to be a regular, even normal event on Wall Street. Sometimes these are caused by simple human error, such as a trader misplacing a zero or using an algorithm intended for a different trade. In other situations, as in the May 2010 flash crash, the causes are more complex. Either way, the underlying conditions for flash crashes remain, making circuit breakers a vital tool for limiting their effects.


A 7% or 13% drop halts trading for 15 minutes. A 20% drop stops trading for the rest of the day.


As Greg Berman, associate director of the SEC’s Office of Analytics and Research, explained, "Circuit breakers don't prevent the initial problems, but they prevent the consequences from being catastrophic." 

Stock trading is one window into what a future of adversarial autonomous systems competing at superhuman speeds might look like. Traders design algorithms to perform a certain function, but the environment they release them into is uncontrolled. Sometimes the algorithms do something different, something that the traders didn’t want them to do. Interactions between algorithms can be unpredictable, in part because firms don’t share their algorithms and trading strategies. When algorithm behavior is too predictable, it can be exploited by spoofing attacks to manipulate algorithms.

The consequence of introducing highly autonomous systems into this complex, chaotic environment can be uncontrolled, runaway processes. Sometimes the harm is isolated, as the Knightmare on Wall Street was. Sometimes the entire ecosystem can collapse, as it did in the flash crash. Normal accident theory, not high reliability, best explains this environment. Accidents are a normal consequence of a complex, tightly coupled system, and automated trading increases the tight coupling of the marketplace by removing the capacity for humans to make trading judgments based on the specific context of the market. The brittleness of algorithms and their inability to flexibly adapt to unanticipated market conditions or manipulation is a major factor in automated trading accidents.

There are parallels between automated stock trading and what autonomous systems might look like in war. Both involve high-speed adversarial interactions in complex, uncontrolled environments. Actual military experience with semi-autonomous and supervised autonomous weapons suggests militaries are equally vulnerable to accidents in such environments.


Militaries have extensive experience with highly automated weapons, including those with semi-autonomous and supervised autonomous modes of operation. The


396 Farrell, “Mini Flash Crashes.”
safety track record of these weapon systems can shed light on militaries’ ability to use autonomous weapons safely. Two case studies are examined: the U.S. Army’s Patriot air and missile defense system and the U.S. Navy’s Aegis weapon system. Both are highly automated air and missile defense systems that have semi-autonomous and supervised-autonomous modes. They are built by the same manufacturer and have roughly equal levels of complexity. However, because they are used by different organizations in different environments, the design of the system, human-machine interaction, role of the operators, training, and procedures are very different between the two weapon systems. An examination of these weapons can aid in understanding militaries’ capacity to use fully autonomous weapons safely.

The Patriot Fraticides

In March 2003, during the Iraq invasion, the Patriot air and missile defense system was deployed to defend coalition forces against attack from Iraqi Scud tactical ballistic missiles. This actual deployment of an operational weapon with both semi-autonomous and supervised autonomous modes provides a real-world case study on how militaries manage these risks.

During the initial invasion, Patriot units were involved in three fratricide incidents, two of which involved Patriot batteries shooting down coalition aircraft. These accidents – both the existence of the accidents and the manner in which they occurred – match what normal accident theory predicts of a complex, tightly-coupled system such as the Patriot.

The first incident occurred on March 23, 2003. A Patriot battery tasked with defending coalition bases in Kuwait shot down a British Tornado GR4A fighter jet descending into Ali Al Salem air base in Kuwait. The accident was a combination of multiple failures that combined to produce a tragic outcome, a classic normal accident.

In addition to targeting tactical ballistic missiles, Patriots have the capability to shoot down anti-radiation missiles, which are homing missiles that can target the Patriot’s radar. Shooting down anti-radiation missiles wasn’t the Patriot operators’ primary job in the Iraq invasion, but they were authorized to engage if the anti-radiation missile appeared to be homing in on their radar, since it would have taken the Patriot battery out of commission. Because of the Tornado’s descending profile as it came in for a landing, the Patriot’s computer tagged the radar signal reflecting off the aircraft as coming from an anti-radiation missile. In the Patriot’s command trailer, the human operators didn’t know that a friendly aircraft was coming in for a landing. Their screen
showed a radar-hunting enemy missile homing in on the Patriot battery. This problem of the radar misclassifying an aircraft as a missile had been identified during operational testing but had not been corrected and was not included in operator training. It was a known limitation of the system, but had not been communicated to the operators. Nevertheless, there were a number of additional safeties that should have prevented this misclassification from escalating into a fatal accident. All of them failed.

The Tornado carried an identification friend or foe (IFF) signal, which was supposed to broadcast a signal to other friendly aircraft and ground radars to let them know their Tornado was friendly and not to fire. But the IFF wasn’t working. The reason why is still unknown. It could be because the Tornado pilots turned it off while over Iraqi territory so as not to give away their position and forgot to turn it back on when returning to Kuwait. It could be because the system simply broke, possibly from a power supply failure. The IFF signal had been tested by maintenance personnel prior to the aircraft taking off, so it should have been functional, but for whatever reason it wasn’t broadcasting. This might not have been a serious problem, but for the fact that the Tornado was targeted by a coalition radar. Even if the Tornado’s IFF had been working, as it turns out, the Patriot wouldn’t have been able to see the signal – the codes for the IFF hadn’t been loaded into the Patriot. The IFF, which was supposed to be a backup safety measure against friendly fire, was doubly broken.

A number of factors complicated the Patriot unit’s situational awareness and hindered their ability to correctly identify the Tornado as a friendly aircraft. The Patriot unit had fallen in on equipment overseas and was using a different, older set of equipment than what they’d trained on. Once in theater, they were detached from their parent battalion and attached to a new battalion whom they hadn’t worked with before. The new battalion was using the newer model equipment, which meant their old equipment (which they weren’t fully trained on in the first place) couldn’t communicate


398 Scharre, Army of None, 144.

399 A UK board of inquiry discounted the possibility that the pilots had intentionally turned off the IFF, although without explaining why. In the absence of other possible explanations, it concluded the IFF system “had a fault.” Ministry of Defence, “Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710.”, 4-5.
with the rest of the battalion. Their systems couldn’t connect to the larger network, depriving them of vital information. All they had was a radio.400

Deprived of the ability to see other radar inputs directly, the 22-year-old Second Lieutenant in charge of the Patriot battery called over the radio to the other Patriot units. Did they see an anti-radiation missile? No one else saw it, but this meant little, since other radars may not have been in a position to see it. The Tornado’s IFF signal, which would have identified the blip on their radar as a friendly aircraft, wasn’t broadcasting.401 There were no reports of coalition aircraft in the area. There was nothing at all to indicate that the blip that appeared on their scopes as an anti-radiation missile might, in fact, be a friendly aircraft.

With only seconds to decide, the 22-year-old Second Lieutenant ordered the Patriot battery to fire, shooting down the aircraft. The missile disappeared from their scope. As far as they knew, it was a successful hit.402 Commanders pieced the fratricide together later after the Tornado never returned home.403

The Army opened an investigation, but the Patriot battery commander remained at her post. There was still a war to fight. The very next night, the same commander shot down an incoming enemy ballistic missile. In all, that same Patriot battery was responsible for 45% of all successful ballistic missile engagements in the war. Later, the investigation cleared the Lieutenant of wrongdoing.

The day after the Tornado shoot down, a different Patriot unit got into a friendly fire engagement with a U.S. F-16 aircraft flying south of Najaf in Iraq. This time, the aircraft shot first. The F-16 fired off a radar-hunting AGM-88 high-speed anti-radiation

400 Ibid.

401 There are actually two relevant IFF modes of operation that might have prevented the Tornado fratricide, Mode 1 and Mode 4. More 4 is the standard encrypted military IFF used by coalition aircraft in theater. It was tested on the Tornado prior to starting engines on the ground and found functional, but there is no evidence it was broadcasting at any point during the flight. The reasons why are unclear. Mode 1 is an unencrypted mode that all coalition aircraft were supposed to be using as a backup. The UK accident investigation report does not specify whether this mode was broadcasting or not. In any case, the Patriot battery did not have the codes for Mode 1, so they couldn’t have received a Mode 1 signal in any case. This is likely because their equipment was not interoperable and so they were not on the network. The Mode 1 codes would have had to have been delivered by hand to be loaded and apparently they were not. Ministry of Defence, “Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710.”


missile. The missile zeroed in on the Patriot’s radar and knocked it out of commission. The Patriot crew was unharmed – a near miss.404

After these incidents, a number of safety measures were put in place to prevent further fratricides. The Patriot has both a manual (semi-autonomous) and auto-fire (supervised autonomous) mode, which can be kept at different settings for different threats. The Patriot can be placed in one setting for tactical ballistic missiles and a different one for anti-radiation missiles. In manual mode, a human is required to approve an engagement before the system will launch. In auto-fire mode, if there is an incoming threat that meets its target parameters, the system will automatically engage the threat on its own.

Because ballistic missiles often afford little reaction time before impact, Patriots sometimes operated in auto-fire mode for tactical ballistic missiles. Given the misidentification of a friendly aircraft as an anti-radiation missile, however, Patriots were required to operate in manual mode for anti-radiation missiles. As an additional safety, systems were now kept in “standby” status so they could track targets, but could not fire without a human bringing the system back to “operate” status. Thus, in order to fire on an anti-radiation missile, two steps were therefore needed: bringing the launchers to operate status and authorizing the system to fire on the target. Ideally, this would prevent another fratricide like the Tornado shootdown. If a system was in auto-fire for ballistic missiles, however, once the system was placed into operation, it would automatically engage ballistic missiles.

In spite of these precautions, a little over a week later on April 2, disaster struck again. A Patriot unit operating north of Kuwait on the road to Baghdad picked up an inbound ballistic missile. Shooting down ballistic missiles was their job. Unlike the anti-radiation missile that the earlier Patriot unit had fired on – which turned out to be a Tornado – there was no evidence to suggest ballistic missiles might be misidentified as aircraft.

What the operators didn’t know – what they could not have known – was that there was no missile. There wasn’t even an aircraft misidentified as a missile. There was nothing. The ballistic missile radar track on their screens was false, a “ghost track” that was likely caused by electromagnetic interference between their radar and another nearby Patriot radar. The Patriot units supporting the U.S. advance north to Baghdad were operating in a non-standard configuration. Units were spread south-to-north along

the main highway to Baghdad in a line instead of the usual widely distributed pattern they would adopt to cover an area. This could have led to electromagnetic interference between the radars and false signals.

But Patriot operators didn’t know this. All they saw was a ballistic missile headed their way. In response, the commander ordered the battery to bring its launchers from “standby” to “operate.” The unit was operating in manual mode for anti-radiation missiles, but auto-fire mode for ballistic missiles. As soon as the launcher became operational, the auto-fire system engaged. Two PAC-3 missiles launched automatically.  

The two PAC-3 missiles steered towards the incoming ballistic missile, or at least to the spot where the ground-based radar told them it should be. The missiles activated their seekers to look for the incoming ballistic missile, but it was a ghost. There was nothing there. Tragically, the missiles’ seekers did find something: a U.S. Navy F/A-18C Hornet fighter jet nearby, which was simply in the wrong place at the wrong time. The F-18 was squawking IFF and showed up on the Patriot’s radar as an aircraft. It didn’t matter. The PAC-3 missiles locked onto the aircraft and zeroed in on him. The pilot saw the missiles coming and called it out over the radio. He took evasive action, but there was nothing he could do. Seconds later, both missiles struck his aircraft, killing him instantly.

Assessing the Patriot’s performance

The Patriot fratricides are an example of the risks of operating complex, highly automated lethal systems. In a strict operational sense, the Patriot units accomplished their mission. Over 60 Patriot fire units were deployed during the initial phase of the war, 40 from the United States and 22 from four coalition nations. Their mission was to protect ground troops from Iraqi ballistic missiles, which they did. Nine Iraqi ballistic missiles were fired at coalition forces; all were successfully engaged by Patriots. No coalition troops were harmed from Iraqi missiles. A Defense Science Board Task Force

---


406 Hess, “Feature: The Patriot’s Fratricide Record.”

407 Ibid.
on the Patriot’s performance concluded that, with respect to missile defense, the Patriot was a “substantial success.”

On the other hand, in addition to these nine successful engagements, Patriots were involved in three fratricides: two incidents in which Patriots shot down friendly aircraft, killing the pilots, and a third incident in which an F-16 fired on a Patriot. Thus, of the 12 total engagements involving Patriots, 25% of them were fratricides, an “unacceptable” fratricide rate according to Army investigators.

The reasons for the Patriot fratricides were a complex mix of human error, improper testing, poor training, and unforeseen interactions on the battlefield. Some problems were known. For example, IFF was well-understood to be an imperfect solution for preventing fratricides. Other problems, such as the potential for the Patriot to misclassify an aircraft as an anti-radiation missile, had been identified during operational testing but had not been corrected and were not included in operator training. Still other problems, such as the potential for electromagnetic interference to cause a false radar track, were novel and unexpected.

One thing that did not happen and was not a cause of the Patriot fratricides is that the Patriot system did not fail, per se. It didn’t break. It didn’t blow a fuse. The system performed its function: it tracked incoming targets and, when authorized, shot them down. Also, in both instances a human was “in the loop” for the engagement. A human had to give the command to fire or at least to bring the launchers to operate. When this lethal, highly automated system was placed in the hands of operators who did not fully understand its capabilities and limitations, however, it turned deadly. Not because the operators were negligent. No one was found to be at fault in either incident. It would be overly simplistic to blame the fratricides on “human error.” Instead, what happened was more insidious. Army investigators determined the Patriot community

---


409 Hawley, “Looking Back at 20 Years of MANPRINT on Patriot.”

410 The Defense Science Board Task Force remarked, “This is not exactly a surprise; this poor performance has been seen in many training exercises. The Task Force remains puzzled as to why this deficiency never garners enough resolve and support to result in a robust fix.” Office of the Under Secretary of Defense For Acquisition, Technology, and Logistics, Report of the Defense Science Board Task Force on Patriot System Performance Report Summary.
had a culture of “trusting the system without question.”

According to Army researchers, the Patriot operators, while nominally in control, exhibited automation bias: an “unwarranted and uncritical trust in automation. In essence, control responsibility is ceded to the machine.”

There may have been a human “in the loop,” but the human operators didn’t question the machine when they should have. They didn’t exercise the kind of judgment Stanislav Petrov did when he questioned the signals his system was giving him. The Patriot operators trusted the machine, and it was wrong.

The Patriot fratricides align with the predictions of normal accident theory. From the perspective of normal accident theory, the fratricides weren’t merely freak occurrences, unlikely to be repeated. Instead, they were a normal consequence of operating a highly lethal, complex, tightly-coupled system. True to normal accidents, the specific chain of events that led to each fratricide was unlikely. Multiple failures happened simultaneously. The Tornado shootdown was caused by the combination of misclassifying the Tornado as an anti-radiation missile, IFF failure, and the human-in-the-loop trusting the automation. If any one of these circumstances had been different, it would not have occurred. The F-18 fratricide similarly stemmed from a combination of failures: electromagnetic interference, human authorization of the engagement, and the F-18 being in the wrong place at the wrong time.

However, simply because these specific combinations of failures were unlikely does not mean that probability of accidents as a whole was low. In fact, given the degree of operational use, the probability of there being some kind of accident was likely quite high. Over sixty Patriot batteries were deployed to Operation Iraqi Freedom, and during the initial phase of the war coalition aircraft flew 41,000 sorties. This means that the number of possible Patriot–aircraft interactions were in the millions. As the Defense Science Board Task Force on the Patriot pointed out, given the sheer number of interactions, “even very-low-probability failures could result in regrettable fratricide

---


413 In fact, Army researchers specifically cited the Three Mile Island incident as having much in common. Hawley, “Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons.”
incidents." The fact that the F-18 and Tornado incidents had different causes lends further credence to the view that normal accidents are lurking below the surface in complex systems, waiting to emerge. The complexities of war may bring these vulnerabilities to the surface.

In fact, one of the most challenges observations from the Patriot fratricide is that while some of the accident causes could have been addressed ahead of time through better testing or training, others could not. War entails uncertainty. Even the best training and operational testing can only approximate the actual conditions of war. Inevitably, soldiers will face wartime conditions they had not expected. The environment, adversary innovation, and simply the chaos, confusion, and violence of war all contribute to unexpected challenges. Loosely coupled systems, such as military organizations comprised of human decision-making, have the flexibility to adapt to novel circumstances in war. Humans can adapt and deviate from guidance if need be in order to accomplish the mission and meet the commander's intent. Automated systems lack this flexibility. This combination of their tight coupling and the near-certainty of unexpected circumstances suggests that accidents are likely inevitable with autonomous weapon systems in wartime environments. This is a pessimistic conclusion from the perspective of managing the risks of autonomous weapons.

The Aegis Combat System

While the Patriot experience during the Iraq invasion in 2003 would appear to align well with the predictions of normal accident theory, the U.S. Navy has operated a similar weapon system with a high degree of reliability for several decades. The Navy’s experience with the Aegis Combat System is an important case study in the feasibility of operating complex, highly automated weapon systems safely. Could high-reliability organizations be a model for how militaries might handle autonomous weapons? In fact, lessons from SUBSAFE and aircraft carrier deck operations have already informed how the Navy operates the human-supervised autonomous weapon, the Aegis combat system.

The Aegis is, in many ways, the archetype for defensive human-supervised autonomous weapon systems. The Navy describes the Aegis as “a centralized, automated, command-and-control (C2) and weapons control system that was designed as a total weapon system, from detection to kill.” It is the electronic brain of the ship’s

---

weapons. The Aegis connects the ship’s advanced radar with its anti-air, anti-surface, and anti-submarine weapon systems and provides a central control interface for sailors. First fielded in 1983, the Aegis has gone through several upgrades and is now used in over 80 U.S. Navy warships.

Aegis is a weapon system of immense complexity. At the core of Aegis is a computer called “Command and Decision,” or C&D, which governs the behavior of the radar and weapons. Command and Decision’s actions are governed by a series of statements – essentially programs or algorithms – that the Navy refers to as “doctrine.” Doctrine statements are instructions for what Command & Decision should do with its radar and weapons in any given situation. Unlike the Patriot circa 2003, however, which had only a handful of different operating modes, Aegis doctrine is almost infinitely customizable.

With respect to weapons engagements, there are essentially four types of human-machine command and control relationships Aegis can use. The most human control is in manual mode, in which engagements against radar “tracks” (objects detected by the radar) must be done directly by the human. Ship commanders can increase the degree of automation in the engagement process by activating one of three types of doctrine: Semi-Auto, Auto SM, and Auto-Special. Semi-Auto, as the term would imply, automates part of the engagement process to generate a firing solution on a radar track, but final decision authority is withheld by the human operator. Auto SM automates more of the engagement process, but a human must still take a positive action before firing. Despite the term, Auto SM still retains a human in the loop. Auto-Special is the only mode where the human is “on the loop” in a supervisory capacity. Once Auto-Special is activated, the Aegis will automatically fire against threats that meet its parameters. The human can intervene to stop the engagement, but no further authorization is needed to fire.

It would be a mistake to think, however, that this means that Aegis can only operate in four modes. In fact, doctrine statements can mix and match these control types against different threats. For example, one doctrine statement could be written to use Auto SM against one type of threat, such as aircraft. Another doctrine statement might authorize Auto-Special against cruise missiles, for which there may be less warning. These doctrine statements can be applied individually or in packages. Captain Pete Galluch, commander of the Aegis Training and Readiness Center which trains all Aegis-qualified officers and enlisted sailors, explained the flexibility of Aegis doctrine. “You can mix and match,” he said. “It’s a very flexible system. … We have the
flexibility so we can do all [doctrine statements] with a push of a button, some with a push of a button, or bring them up individually."\textsuperscript{415}

This makes Aegis less like a finished product with a few different modes and more like a customizable system that can be tailored for each mission. Galluch explained that the ship’s doctrine review board, consisting of the officers and senior enlisted personnel who work on Aegis, will begin the process of writing doctrine months before deployment. They will factor their anticipation missions, intelligence assessments, and information on the region for the upcoming deployment into their recommendations, which go to the ship’s captain for approval. The result is a series of doctrine statements, individually and in packages, that the captain can activate as needed during deployment. “If you have your doctrine statements built and tested,” Galluch said, the time to “bring them up is seconds.”\textsuperscript{416}

Doctrine statements are typically grouped into two general categories: non-saturation and saturation. Non-saturation doctrine is used when there is time to carefully evaluate each potential threat. Saturation doctrine is needed if the ship gets into a combat situation where the number of inbound threats could overwhelm the ability of operators to respond. “If World War III starts and people start throwing a lot of stuff at me,” Galluch said, “I will have grouped my doctrine together so that it’s a one-push button that activates all of them. And what we’ve done is we’ve tested and we’ve looked at how they overlap each other and what the effects are going to be and make sure that we’re getting the defense of the ship that we expect.” This is where something like Auto-Special comes into play, in a “kill or be killed” scenario, Galluch said.\textsuperscript{417}

It’s not enough to build the doctrine, though. Extensive testing goes into ensuring that it works properly. Once the ship arrives in theater, the first thing the crew does is test the weapons doctrine to see if there is anything in the environment that might cause it to fire in peacetime, which would not be good. This is done safely by enabling a hardware-level cutout called the Fire Inhibit Switch, or FIS. The FIS includes a key that must be inserted in order for any of the ship’s weapons to fire. When the FIS key is inserted a red light comes on; when it is turned to the right, the light turns green, meaning the weapons are live and ready to fire. When the FIS is red – or removed entirely – the ship’s weapons are disabled at the hardware level. As Galluch put it,

\textsuperscript{415} Captain Peter Galluch, Commander of the Aegis Training and Readiness Center, U.S. Navy interview by author, Dahlgren, VA, July 15, 2016.

\textsuperscript{416} Ibid.

\textsuperscript{417} Ibid.
“there is no voltage that can be applied to light the wick and let the rocket fly out.”

By keeping the FIS red or removing the key, the ship’s crew can test Aegis doctrine statements safely without any risk of inadvertent firing.

Establishing the doctrine and activating it is the sole responsibility of the ship’s captain. Doctrine is more than just a set of programs. It is the embodiment of the captain’s intent for the warship. “Absolutely, it’s automated, but there’s so much human interface with what gets automated and how we apply that automation,” Galluch said. Aegis doctrine is a way for the captain to pre-delegate his or her decision-making against certain threats.

The way automation is used in the Aegis is very different than how it was used in the Patriot circa 2003. Patriot operators sitting at the consoles in 2003 were essentially trusting in the automation. They had a handful of operational modes they could activate, but the operators themselves didn’t write the rules for how the automation would function in those modes. Those rules were written years beforehand. The Aegis, by contrast, can be customized and tailored to the specific operating environment. But the differences run deeper than merely having more options – the whole philosophy of automation is different. With Aegis, the automation is used to capture the ship captain’s intent. In Patriot, the automation embodies the intent of the designers and testers. The actual operators of the system may not even fully understand the designers’ intent that went into crafting the rules. The automation in Patriot is largely intended to replace warfighters’ decision-making. In Aegis, the automation is used to capture warfighters’ decision-making.

Another key difference is where decision authority rests. Only the captain of the ship has the authority to activate Aegis weapons doctrine. The captain can pre-delegate that authority to the tactical action officer on watch, but it must be in writing as part of official orders. This means the decision maker’s experience level for Aegis operations is radically different from Patriot. When Captain Galluch took command of the USS Ramage, he had 18 years of experience and had served on three prior Aegis ships. By contrast, the person who made the call on the first Patriot fratricide was a 22-year old Second Lieutenant fresh out of training.

---

418 Ibid.
419 Ibid.
420 Ibid.
Galluch explained the seriousness of activating Aegis doctrine. “You’re never driving around with any kind of weapons doctrine activated” unless you expect to get into a fight, he explained. Even on manual mode, it is possible to launch a missile in seconds. And if need be, doctrine can be activated quickly. “I’ve made more Gulf deployments than I care to,” he said. “I’m very comfortable with driving around for months at a time with no active doctrine but making damn sure that I have it set up and tested and ready to go if I need to.” Because there can be situations that call for that level of automation. “You can get a missile fired pretty quickly, so why don’t you do everything manually?” Galluch explained: “My view is that [manual control] works well if it’s one or two missiles or threats. But if you’re controlling fighters, you’re doing a running gun battle with small patrol boats, you’re launching your helicopter … and you’ve got a bunch of cruise missiles coming in from different angles. You know, the watch is pretty small. It’s ten or twelve people. So, there’s not that many people … You can miss things coming in. That’s where I get to the whole concept of saturation vs. normal. You want the man in the loop as much as possible, but there comes a time when you can get overwhelmed.”

At the core of the Aegis philosophy is one of human control over engagements, even when doctrine is activated. What varies is the form of human control. In Auto-Special doctrine, the engagement decision is delegated to Aegis’ Command & Decision computer, but the human intent is still there. The goal is always to ensure “there is a conscious decision to fire a weapon,” Galluch said. That doesn’t mean that accidents can’t happen. In fact, it is the constant preoccupation with the potential for accidents that helps prevent them. Galluch and others understand that, with doctrine activated, mishaps can happen. That’s precisely why tight control is kept over the weapon. “[Ship commanding officers] are constantly balancing readiness condition to fire the weapon versus a chance for inadvertent firing,” he explained.

“Roll green”

At the Aegis simulation center, Galluch and his team demonstrated a series of mock engagements. Galluch stood in as the ship’s commanding officer and had Aegis-qualified sailors sitting at the same terminals doing the same jobs they would on a real ship.

---

421 Ibid.

422 Ibid.
First, Galluch ordered the sailors to demonstrate a shot in manual operation. They put a simulated radar track on the screen and Galluch ordered them to target the track. They began working a firing solution, with the three sailors calmly but crisply reporting when they had completed each step in the process. Once the firing solution was ready, Galluch ordered the tactical action officer to roll his FIS key to green. (The command, as reported on the Navy’s website, is “roll FIS green.”)\(^{223}\) The tactical action officer turned the key and the light switched from red to green. Then Galluch gave the order to fire. A sailor pressed the button to fire and called out that the missile was away. On a large screen in front of us, the radar showed the outbound missile racing towards the track.

The whole process had been exceptionally fast – under a minute. The threat had been identified, a decision made, and a missile launched well under a minute, and that was in manual mode.

They did it again in Semi-Auto mode, now with doctrine activated. The FIS key was back at red, the tactical action officer having turned it back right after the missile was launched. Galluch ordered them to activate Semi-Auto doctrine. Then they brought up another track to target. This time, Aegis’ Command & Decision computer generated part of the firing solution automatically. This shortened the time to fire by more than half.

They rolled FIS red, activated Auto SM doctrine, and put up a new track. Roll FIS green. Fire.

Finally, they brought up Auto-Special doctrine. This was it. This was the most autonomous mode with the human removed from the loop. The sailors were merely observers now; they didn’t need to take any action for the system to fire. Except … The FIS key was in, but it was turned to red. Auto-Special doctrine was enabled, but there was still a hardware-level cutout in place. There was not even any voltage applied to the weapons. Nothing could fire until the tactical action officer rolled his key green.

The track for a simulated threat came up on the screen and Galluch ordered them to roll FIS green. It was only a handful of heartbeats before a sailor announced the missiles were away. That’s all it took for Command & Decision to target the track and fire.

The speed with which the Aegis could targeting incoming threats was impressive, but the degree of human control over the system was even more so. Even on

Auto-Special, and they had their hand literally on the key that, with a flick of the wrist, disabled firing. And as soon as the missile was away, the tactical action officer rolled FIS red again. The automation was powerful and the Aegis operators employed it when necessary, but that didn’t mean they were surrendering their human decision-making to the machine.

To further drive the point home, Galluch had them demonstrate one final shot. With Auto-Special doctrine enabled, they rolled FIS green and let Command & Decision take its shot. But then after the missile was away, Galluch ordered them to abort the missile. They pushed a button and a few seconds later the simulated missile disappeared from our radar, having been destroyed mid-flight. Even in the case of Auto-Special, even after the missile had been launched, they still had the ability to reassert human control over the engagement.

The Aegis community has reason to be careful. In 1988, the USS Vincennes, an Aegis-equipped cruiser, shot down commercial airliner Iran Air Flight 655 over the Persian Gulf, killing all 290 people on board. Galluch described the Vincennes incident as a “terrible, painful lesson” and what the Aegis community learned to prevent future tragedies.424

424 Galluch, interview.

Iran were high afterward. U.S. ships in the area were on high alert to the ever-present threat from Iranian mines and rocket-armed boats.

On the morning of July 3, 1988, the U.S. warships USS *Vincennes* and USS *Montgomery* were escorting tankers through the Strait of Hormuz when they came into contact with Iranian fast boats. The *Vincennes* dispatched its helicopter to monitor the Iranian boats and the helicopter came under fire. *Vincennes* and *Montgomery* responded, pursuing the Iranian boats into Iranian territorial waters and opening fire.

While *Vincennes* was in the midst of a gun battle with the Iranian boats, two aircraft took off in close sequence from Iran’s nearby Bandar Abbas airport. Bandar Abbas was a dual-use airport, servicing both Iranian commercial and military flights. One aircraft was a commercial airliner, Iran Air Flight 655. The other was an Iranian F-14 fighter. For whatever reason, in the minds of the sailors in the *Vincennes’* combat information center, the tracks of the two aircraft on their radar screens became confused. The Iranian F-14 veered away but Iran Air 655 flew along its normal commercial route, which happened to be directly towards the *Vincennes*. Even though the commercial jet was squawking IFF, the *Vincennes* captain and crew became convinced, incorrectly, that the radar track headed towards their position was an Iranian F-14 fighter.

As the aircraft approached, *Vincennes* issued multiple warnings on military and civilian frequencies. There was no response. Believing the Iranians were choosing to escalate the engagement by sending a fighter and that his ship was under threat, the *Vincennes’* captain gave the order to fire. Iran Air 655 was shot down with no survivors.426

The USS *Vincennes* incident and the Patriot fratricides sit as two opposite cases on the scales of automation vs. human control. In the case of the Patriot fratricides, humans trusted in the automation too much. The *Vincennes* incident was caused by human error and more automation might have helped. Iran Air 655 was squawking IFF, and well-crafted Aegis doctrine should not have fired. The *Vincennes* crew faced a different kind of saturation problem. They weren’t overwhelmed with too many missiles, but they were overwhelmed with too much information: the running gun battle with Iranian boats and tracking an F-14 and a commercial airliner launching in close succession from a nearby airport. In this information-saturated environment, the crew

---

426 This account comes from an ABC special on the *Vincennes* incident which relies on first-hand interviews and video and audio recordings from the *Vincennes* during the incident. ABC Four Corners, “Shooting Down of Iran Air 655,” 2000, [https://www.youtube.com/watch?v=OnkJ3ZVME](https://www.youtube.com/watch?v=OnkJ3ZVME).
missed important details they should have noticed and made poor decisions with grave consequences. Automation, by contrast, wouldn’t have gotten overwhelmed by the amount of information. Just as automation could help shoot down incoming missiles in a saturation scenario, it could also help *not fire* at the wrong targets in an information-overloaded environment.

*The challenge of achieving high reliability*

The Aegis community has learned from the *Vincennes* incident, the Patriot fratricides, and years of experience to refine their operating procedures, doctrine, and software to the point where they are able to operate a very complex weapon system with low accidents. In the nearly 30 years since *Vincennes*, there has not been another similar incident, even with Aegis ships deployed continuously around the world.

The Navy’s track record with Aegis shows that high-reliability operation of complex, hazardous systems is possible, but it doesn’t come from testing alone. The human operators are not passive bystanders in the Aegis’ operation, trusting blindly in the automation. They are active participants at every stage. They program the system’s operational parameters, constantly monitor its modes of operation, supervise its actions in real-time, and maintain tight control over weapons release authority. The Aegis culture is 180 degrees from the “unwarranted and uncritical trust in automation” that Army researchers found in the Patriot community in 2003.427

After the Patriot fratricides, the Army launched the Patriot Vigilance Project, a three-year post-mortem assessment to better understand what went wrong and improve training, doctrine, and system design to ensure it didn’t happen again. Dr. John Hawley is an engineering psychologist who led the project and spoke frankly about the challenges in implementing those changes. He said that there are examples of communities that have been able to manage high-risk technologies with very low accident rates, but high reliability is not easy to achieve. The Navy “spent a lot of money looking into … how you more effectively use a system like Aegis so that you don’t make the kinds of mistakes that led to the [*Vincennes* incident].”428 This training is costly and time-consuming, and in practice there are bureaucratic and cultural obstacles that may prevent military organizations from investing this amount of effort. Hawley explained that Patriot commanders are evaluated based on how many trained

427 Hawley, “Not by Widgets Alone.”

428 John Hawley, engineering psychologist, Army Research Laboratory, interview by author, telephone, December 5, 2016.
crews they keep ready. “If you make the [training] situation too demanding, then you could start putting yourself in the situation where you’re not meeting those requirements.” Hawley argued that Army Patriot operators train in a “sham environment” that doesn’t accurately simulate the rigors of real-world combat. As a result, he said “the Army deceives itself about how good their people really are. … It would be easy to believe you’re good at this, but that’s only because you’ve been able to handle the relatively non-demanding scenarios that they throw at you.” It may seem that militaries have an incentive to make training as realistic as possible, and to a certain extent that’s true, but there are limits to how much time and money can be applied. Militaries might not realize their training is ineffective until a war occurs, at which point it may be too late.

Hawley explained that the Aegis community was partially protected from this problem because they use their system day-in and day-out on ships operating around the globe. Aegis operators get “consistent objective feedback from your environment on how well you’re doing,” preventing this kind of self-deception. The Army’s peacetime operating environment for the Patriot, on the other hand, is not as intense, Hawley said. “Even when the Army guys are deployed, I don’t think that the quality of their experience with the system is quite the same. They’re theoretically hot, but they’re really not doing much of anything, other than just monitoring their scopes.” Leadership is also a vital factor. “Navy brass in the Aegis community are absolutely paranoid” about another Vincennes incident, Hawley said.

The bottom line is that high reliability not easy to achieve. It requires frequent experience under real-world operating conditions and a major investment in time and money. Safety must be an overriding priority for leaders, who often have other demands they must meet. U.S. Navy submariners, aircraft carrier deck operators, and Aegis weapon system operators are very specific military communities that meet these conditions. Military organizations in general do not. Hawley was pessimistic about the ability of the U.S. Army to safely operate a system like the Patriot, saying it was “too sloppy an organization to … insist upon the kinds of rigor that these systems require.”

This is a disappointing conclusion, because the U.S. Army is one of the most professional military organizations in history. Hawley was even more pessimistic about
other nations. “Judging from history and the Russian army’s willingness to tolerate casualties and attitude about fratricide ... I would expect that ... they would tilt the scale very much in the direction of lethality and operational effectiveness and away from necessarily safe use.” Practice would appear to bear this out. The accident rate with Soviet/Russian submarines is far higher than U.S. submarines.

_Evaluating the Aegis and Patriot_

The Aegis and Patriot are valuable case studies because they have the same basic mission (air and missile defense), are used by the same country (United States), built by the same developer (Raytheon), developed over roughly the same time frame (Aegis first fielded in 1983 and Patriot in 1984), and have the same high-level command-and-control concept (semi-autonomous most of the time with the ability to go to supervised autonomous if need be for speed). Yet in other ways the systems are profoundly different.

Control over the Aegis weapon system and activating Aegis doctrine is held at a much higher level in the Navy than control over the Patriot the Army. This means that a more experienced officer is calling the shots for the Aegis, one with multiple prior deployments. For the Army, the Patriot commander deciding whether to authorize an engagement could be on his or her first deployment.

Additionally, the Aegis is more customizable. Aegis doctrine, the ruleset for engagements, is tailored for each mission. While in both systems the automation is the instantiation of human intention, the key difference is which humans’ intent is captured in the automation. For the Aegis, the customizability of the automation means that the automation is capturing the intent of the Navy ship commander. For the Patriot, it is the intent of the system’s designers and engineers. The Patriot operator is merely choosing between a few limited modes. It is fair to say that in the Patriot, the automation is replacing the human operator’s decision-making, while in the Aegis, the automation is intended to capture the human operator’s intentions. In short, the Aegis has a much closer human-machine relationship and integration with the warfighter.

---

432 Ibid.
Comparison of Patriot vs. Aegis Missile Defense Systems

<table>
<thead>
<tr>
<th></th>
<th>Patriot</th>
<th>Aegis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mission</td>
<td>Air and missile defense</td>
<td>Air and missile defense</td>
</tr>
<tr>
<td>Environment</td>
<td>Land</td>
<td>Sea</td>
</tr>
<tr>
<td>Originating country</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>Service</td>
<td>Army</td>
<td>Navy</td>
</tr>
<tr>
<td>Human-machine relationship</td>
<td>Semi-autonomous with</td>
<td>Semi-autonomous with the</td>
</tr>
<tr>
<td></td>
<td>the option to go to</td>
<td>option to go to supervised</td>
</tr>
<tr>
<td></td>
<td>supervised autonomous</td>
<td>autonomous</td>
</tr>
<tr>
<td>Modes of operation</td>
<td>Limited modes</td>
<td>Customizable</td>
</tr>
<tr>
<td>Level of human decision-maker</td>
<td>O-1 (Second lieutenant)</td>
<td>O-5 (Commander)</td>
</tr>
<tr>
<td>Can abort missile in flight</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Perhaps one of the most significant differences, however, is the fact that the Aegis is used routinely in peacetime. Auto-special doctrine is not normally activated, but Navy ships around the world use the Aegis weapon system in real-world environments on a day-to-day basis all the time. This means that the operators can build up a level of trust about how the system will interact with normal environmental conditions. This is an important factor in improving safety. While the Patriot had a very high fratricide rate in the Iraq invasion, with 3 of 12 Patriot-related engagements as fratricides (25%), there is every reason to believe that if Patriot operations had continued for an extended period of time that the accident rate would have declined. Warfighters – both Patriot operators and allied pilots – would have changed their tactics, techniques, and procedures as a result of these fratricides, learning from these incidents. Indeed, Patriot operations did change significantly over the course of the Iraq invasion following the fratricides. Conversely, it is entirely reasonable to expect that if the Aegis were used in wartime, that accidents may occur. While the Navy has used Aegis to a high degree of reliability in peacetime for several decades, these experiences may not translate well to a wartime environment. Wartime operations may introduce novel circumstances for Aegis operators, as it did in the case of the Patriot in 2003. These could result in fratricides or accidents involving civilian aircraft, as occurred with the USS Vincennes in 1988.
From the Patriot and Aegis case studies, one can draw several conclusions. First, the Patriot fratricides appear to align well with the predictions of normal accident theory. This suggests that highly automated systems like the Patriot or Aegis ought to be vulnerable to the kinds of accidents that normal accident theory predicts. Second, the Aegis operational track record post-Vincennes suggests that high reliability can be achieved provided certain conditions are met. These conditions include: (1) routine, day-to-day operations so that human operators can understand the environment and how it interacts with their system and make adjustments as needed; (2) a high degree of human involvement and control over the system.

Even if a high degree of human involvement is needed to achieve high reliability, incidents like the USS Vincennes shoot-down of Iran Air 655 highlight the fact that humans also make mistakes. In some cases, more automation may be beneficial. As militaries introduce more automation into weapon systems, does the likelihood or consequence of accidents increase or diminish?

**Automation as a Source of Both Reliability and Failure**

Automation plays a mixed role in accidents. Sometimes the brittleness and inflexibility of automation can cause accidents. In other situations, automation can help reduce the probability of accidents or mitigate their damage. At Fukushima Daiichi, automated safeties scrammed the reactor and brought backup generators online. Is more automation a good or bad thing?

We have two intuitions when it comes to autonomous systems, intuitions that come partly from science fiction but also from our everyday experiences with phones, computers, cars, and myriad other computerized devices.

The first intuition is that autonomous systems are reliable and introduce greater precision. Just as autopilots have improved air travel safety, automation can also improve safety and reliability in many other domains. Humans are terrible drivers, for example, killing more than 30,000 people a year in the United States alone (roughly the equivalent of a 9/11 attack every month).\(^{433}\) Even without fully autonomous cars, more

advanced vehicle autopilots that allow cars to drive themselves under most conditions could dramatically improve safety and save lives.\textsuperscript{434}

However, we have another instinct when it comes to autonomous systems, and that is one of robots run amok, autonomous systems that slip out of human control and result in disastrous outcomes. These fears are fed to us through a steady diet of dystopian science fiction stories in which murderous AIs turn on humans, from \textit{2001: A Space Odyssey}’s HAL 9000 to \textit{Ex Machina}’s Ava. But these intuitions also come from our everyday experiences with simple automated devices. Anyone who has ever been frustrated with an automated telephone call support helpline, an alarm clock mistakenly set to “p.m.” instead of “a.m.,” or any of the countless frustrations that come with interacting with computers has experienced the problem of “brittleness” that plagues automated systems. Autonomous systems will do precisely what they are programmed to do, and it is this quality that makes them both reliable and maddening, depending on whether what they were programmed to do was the right thing at that point in time.

Both of these intuitions are correct. If they are designed, tested, and used the right way, autonomous systems can often perform tasks far better than humans. They can be faster, more reliable, and more precise. However, if they are placed into situations for which they were not designed, if they aren’t fully tested, if operators aren’t properly trained, or if the environment changes, then autonomous systems can fail. When they do fail, they often fail badly. Unlike humans, autonomous systems lack the ability to step outside their instructions and employ “common sense” to adapt to the situation at hand. In this sense, they are “brittle.” Humans can adjust on the fly, but autonomous systems will mindlessly follow their programming, even if it no longer makes sense.

This problem of brittleness was highlighted during a telling moment in the 2011 \textit{Jeopardy! Challenge} in which IBM’s Watson AI took on human Jeopardy champions Ken Jennings and Brad Rutter. Towards the end of the first game, Watson momentarily stumbled in its rout of Jennings and Rutter in response to a clue in the “Name the Decade” category. The clue was, “The first modern crossword is published and Oreo cookies are introduced.” Jennings rang in first with the answer, “What are the 20s?” Wrong, said host Alex Trebek. Immediately afterward, Watson rang in and gave the

\textsuperscript{434} For example “Intelligent Drive,” Mercedes-Benz, https://www.mbusa.com/mercedes/technology/videos/detail/title-safety/videoId-fc0835ab8d127410VgnVCM100000ccee1e35RCRD.
same answer, “What is 1920s?” A befuddled Trebek testily replied, “No, Ken said that.”

Any human would have known after listening to Jennings’ response that “What is 1920s?” was the incorrect answer. Watson hadn’t been programmed to listen to other contestants’ wrong answers and adjust accordingly, however. Processing Jennings’ answer was outside of the bounds of Watson’s design. Watson was superb at answering Jeopardy questions under most conditions, but its design was brittle. When an atypical event occurred that Watson’s design didn’t account for, such as Ken Jennings getting a wrong answer, Watson couldn’t adapt on the fly. As a result, Watson’s performance suddenly plummeted from superhuman to super-dumb.

Brittleness can be managed when the person using an autonomous system understands the boundaries of the system – what it can and cannot do. The user can either steer the system away from situations outside the bounds of its design or knowingly account for and accept the risks of failure. In this case, Watson’s designers understood this limitation. They just didn’t think that the ability to learn from other contestants’ wrong answers would be important. “We just didn't think it would ever happen,” one of Watson’s programmers said afterward. Watson’s programmers were probably right to discount the importance of this ability. The momentary stumble proved inconsequential. Watson handily defeated its human counterparts.

Problems can arise when human users don’t anticipate these moments of brittleness. This was the case with the Patriot fratricides. The system had vulnerabilities – misclassifying an anti-radiation missile as an aircraft, IFF failures, and electromagnetic interference causing “ghost track” ballistic missiles – that the human operators were either unaware of or didn’t sufficiently account for. As a result, the human user’s expectations and the system’s actual behavior were not in alignment. The operators thought the system was targeting missiles when it was actually targeting aircraft.

---

435 Bin Kenney, “Jeopardy!—The IBM Challenge (Day 1—February 14),” video, [https://www.youtube.com/watch?v=i-vMW_Ce51w](https://www.youtube.com/watch?v=i-vMW_Ce51w).


437 Ibid.
Autonomy and risk

Professor William Kennedy of George Mason University has a unique background to help understand the role of automation in safety. He spent 30 years in the Navy (active and reserve) on nuclear missile submarines, combined with 25 years working for the Nuclear Regulatory commission and the Department of Energy on nuclear reactor safety. To top it off, Kennedy has a PhD in information technology with a focus on artificial intelligence. Kennedy said, “A significant message for the Nuclear Regulatory Commission from Three Mile Island was that humans were not omnipotent. The solution prior to Three Mile Island was that every time there was a design weakness or a feature that needed to be processed was to give the operator another gauge, another switch, another valve to operate remotely from the control room and everything would be fine. And Three Mile Island demonstrated that humans make mistakes. … We got to the point where we had over 2,000 alarms in the control rooms, a wall of procedures for each individual alarm. And Three Mile Island said that alarms almost never occur individually.” This was an unmanageable level of complexity for any human operator to absorb, Kennedy explained.438

Following Three Mile Island, more automation was introduced to manage some of these processes. Automation can be beneficial in the right circumstances. “The automated systems, as they are currently designed and built, may be more reliable than humans for planned emergencies, or known emergencies. … If we can study it in advance and lay out all of the possibilities and in our nice quiet offices consider all the ways things can behave, we can build that into a system and it can reliably do what we say. But we don’t always know what things are possible. … Machines can repeatedly, quite reliably, do planned actions. … But having the human there provides for beyond design basis accidents or events.” Automation can help for situations that can be predicted, but humans are needed to manage novel situations to compensate for automation’s brittleness. It is overly simplistic to count solely on humans or automation, Kennedy explained. “Both sides have strengths and weaknesses. They need to work together, at the moment, to provide the most reliable system.”439

One of the ways to compensate for the brittle nature of automated systems is to retain tight control over their operation. If the system fails, humans can rapidly

438 William Kennedy, Associate Professor in the Center for Social Complexity and the Department of Computational and Data Science, George Mason University, interview by author, Fairfax, VA, December 8, 2015.

439 Ibid.
interact to correct it or halt its operation. Tighter human control reduces the autonomy, or freedom, of the machine. In this case, the automation can play a valuable role for known emergencies but the human is tightly integrated and can intervene if needed in novel situations. This type of close human-machine integration that leverages the best of both forms of cognition is only possible, however, in some types of systems.

Immediate human intervention is possible for semi-autonomous and human-supervised systems. Just because humans can intervene, however, doesn’t mean they always do when they should. In the case of the Patriot fratricides, humans were “in the loop,” but they didn’t sufficiently question the automation. Humans can’t act as an independent fail-safe if they cede their judgment to the machine.

Effective human intervention may be even more challenging in supervised autonomous systems, where the system does not pause to wait for human input. The human’s ability to actually regain control of the system in real-time depends heavily on the speed of operations, the amount of information available to the human, and any time delays between the human’s actions and the system’s response. Giving a driver the ability to grab the wheel of an autonomous vehicle traveling at highway speeds in dense traffic, for example, is merely the illusion of control, particularly if the operator is not paying attention. For example, this appears to have been the case in a 2016 fatality involving a Tesla Model S driven while on autopilot.\(^440\)

For fully autonomous systems, the human is out of the loop and cannot intervene at all, at least for some period of time. This means that if the system fails or the context changes, it will keep doing what it was programmed to do, no matter how senseless the action might be. The result could be a runaway autonomous process beyond human control with no ability to halt or correct it.

This danger of autonomous systems is best illustrated not with a science fiction story, but with a Disney cartoon. The Sorcerer’s Apprentice is an animated short in Disney’s 1940 film Fantasia.\(^441\) In the story, which is an adaptation of an eighteenth-century poem by Goethe, Mickey Mouse plays the apprentice of an old sorcerer. When the sorcerer leaves for the day, Mickey decides to take it upon himself to put his novice

---


magic to good use to automate his chores. Mickey dons the sorcerer’s hat and enchants a broomstick, bringing it to life. At Mickey’s beckon, the broomstick leaps forward, sprouts arms, and picks up two pails to carry water – a chore Mickey is supposed to be doing. The broomstick dutifully follows Mickey as he shows it how to draw water from a well and carry the now sloshing-over pails to a cistern and fill it with water. In short order, Mickey is relaxing in the sorcerer’s chair, his feet up on the table, his chores now automated.

Mickey watches the broomstick for a while as it fetches more loads of water, but soon falls asleep. While Mickey snoozes, the autonomous process he has set up keeps running its course. The cistern fills with water and overflows. The task is complete, but the broomstick doesn’t know this. No one told it to stop. Mickey awakes to find the room flooded, his chair floating like a piece of driftwood at sea. Mickey commands the broomstick to halt, but it doesn’t comply. Finally, in a moment of desperation, Mickey snatches an axe from the wall and attacks the broomstick, splintering it to pieces.

As Mickey catches his breath, the splinters reanimate, each piece growing into another broomstick with arms and a pail in each hand. The horde of broomsticks march forth to carry even more water, an army of rogue autonomous agents out of control. Even as the water level rises, filling the whole room, they senselessly continue their task of hauling water. Finally, the madness is stopped only by the return of the sorcerer himself. With a wave of his arms, the sorcerer parts the water and returns the broomstick to its original, inanimate form.

With the original German poem written in 1797, *The Sorcerer’s Apprentice* may be the first example of autonomy displacing jobs. It also shows the danger of automation. The autonomous broomsticks are brittle agents. They know only their task. They don’t understand the context for use. They don’t know when the task is complete and they should stop hauling water. In the original Goethe poem, the apprentice forgets the word to command the broomstick to cease. He’s not properly trained – or at least he wasn’t paying attention to his training – for the autonomous agents he’s using.442

*The Sorcerer’s Apprentice* illustrates the risk in using autonomous agents. The danger in delegating a task to an autonomous system is that it may not perform the task in the manner we want. This could occur for a variety of reasons: malfunction, user error, unanticipated interactions with the environment, or hacking. In the case of

Mickey’s problem, the “software” (instructions) that he bewitched the broomstick with were flawed because he didn’t specify when to stop. Overfilling the cistern might have been only a minor annoyance if it had happened once, however. A semi-autonomous process that paused for Mickey’s authorization after each trip to the well would have been far safer. Having a human “in the loop” would have mitigated the danger from the faulty software design. Making the broomstick fully autonomous without a human in the loop wasn’t the cause of the failure, but it did dramatically increase the consequences if something went wrong.

Because of this potential to have a runaway process, fully autonomous systems are inherently more hazardous than semi-autonomous ones that keep a human in the loop. A human in the loop acts is a natural fail-safe. Like a firebreak in the trees to stop wildfires or a circuit breaker in an electrical circuit, a human in the loop can catch harmful events before they cascade into disaster. Without that human circuit breaker, things can spiral out of control. Autonomous systems will continue mindlessly performing their task, regardless of whether it makes sense anymore.

Putting an autonomous system into operation means accepting the risk that it may perform its task inappropriately, and that it might continue doing so for some period of time. Fully autonomous systems are not necessarily more likely to fail than semi-autonomous or supervised autonomous ones, but if they do, the consequences could be more severe. The damage potential is the amount of harm it could cause before the next opportunity for human intervention. For powerful autonomous systems, the damage potential could be quite high.

_Trust, but verify_

Activating an autonomous system is an act of trust. The user trusts that the system will function in the manner that he or she expects. Trust isn’t blind faith, however. As the Patriot fratricides demonstrated, too much trust can be just as dangerous as too little. Human users need to trust the system just the right amount, not too much and not too little. In order to do this, humans need to understand both the capabilities and limitations of the system. A rigorous testing regime can help designers and operators better understand how the system performs under realistic conditions.

The problem is that, even with simulations that test millions of scenarios, fully testing all of the possible scenarios a complex autonomous system might encounter is effectively impossible. There are simply too many possible interactions between the system and its environment and even within the system itself. Mickey should have been able to anticipate the cistern overflowing, but some real-world problems cannot be
anticipated. The game of Go has more possible positions than atoms in the universe, and the real world is far more complex than Go. A 2015 Air Force report on autonomy bemoaned the problem:

Traditional methods … fail to address the complexities associated with autonomy software … There are simply too many possible states and combination of states to be able to exhaustively test each one.\textsuperscript{443}

In addition to the sheer numerical problem of evaluating all possible combinations, testing is also limited by the testers’ imagination. In games like chess or Go, the set of possible actions is limited. In the real world, however, autonomous systems will encounter any number of novel situations: new kinds of human error, unexpected environmental conditions, or creative actions by adversaries looking to exploit vulnerabilities. If these scenarios can’t be anticipated, they can’t be tested.

Testing is vital to building confidence in how autonomous systems will behave in real world environments, but no amount of testing can entirely eliminate the potential for unanticipated behaviors. When playing games like Jeopardy, chess, or Go, surprising behaviors may be tolerable, even interesting flukes. When operating high-risk automated systems where life or death is at stake, unexpected actions can lead to tragic accidents, such as the Patriot fratricides.

\textit{Delegating Autonomy}

To use fully autonomous weapons is to accept the risk of the potential harm they may cause if they fail. John Borrie of UNIDIR said, “I think that we’re being overly optimistic if we think that we’re not going to see problems of system accidents” in autonomous weapons.\textsuperscript{444} Army engineering psychologist John Hawley agreed: “If you’re going to turn these things loose, whether it be Patriot, whether it be Aegis, whether it be some type of totally unmanned system with the ability to kill, you have to be psychologically prepared to accept the fact that sometimes incidents will happen.”\textsuperscript{445} Charles Perrow, the father of normal accident theory, made a similar conclusion about complex systems in general:


\textsuperscript{444} Borrie, interview.

\textsuperscript{445} Hawley, interview.
Even with our improved knowledge, accidents and, thus, potential catastrophes are inevitable in complex, tightly coupled systems with lethal possibilities. We should try harder to reduce failures – and that will help a great deal – but for some systems it will not be enough. … We must live and die with their risks, shut them down, or radically redesign them. 446

If we are to use autonomous weapons, we must accept their risks. But how hazardous are autonomous weapons? Even if autonomous weapons were to occasionally fail, so what? All weapons are dangerous. Indeed, that is the point. War entails violence. Weapons that are designed to be dangerous to the enemy can also be dangerous to the user if they slip out of control. Even a knife wielded improperly can slip and cut its user. Most modern weapons, regardless of their level of autonomy, are complex systems. Accidents will happen, and sometimes these accidents will result in fratricide or civilian casualties. What makes autonomous weapons any different?

The key difference between semi-, human-supervised, and fully autonomous weapons is amount of damage the system can cause until the next opportunity for a human to intervene. In semi-autonomous or supervised autonomous systems, such as Aegis, the human is a natural fail-safe against accidents, a circuit breaker if things go wrong. The human can step outside of the rigid rules of the system and exercise judgment. Taking the human out of the loop reduces slack and increases the coupling of the system. In fully autonomous weapons, there is no human to intervene and halt the system’s operation. A failure that might cause a single unfortunate incident with a semi-autonomous weapon could cause far greater damage if it occurred in a fully autonomous weapon.

The runaway gun

A simple malfunction in an automatic weapon – a machine gun – provides an analogy for the danger with autonomous weapons. When functioning properly, a machine gun continues firing so long as the trigger remains held down. Once the trigger is released, a small metal device called a sear springs into place to stop the operating rod within the weapon from moving, halting the automatic firing process. Over time,

446 Perrow, Normal Accidents, 354.
however, the sear can become worn down. If the sear becomes so worn down that it fails to stop the operating rod, the machine gun will continue firing even when the trigger is released. The gun will keep firing on its own until it exhausts its ammunition. This malfunction is called a runaway gun.

Runaway guns are serious business. The machine gunner has let go of the trigger, but the gun continues firing. The firing process is now fully automatic with no way to directly halt it. The only way to stop a runaway gun is to break the links on the ammunition belt feeding into the weapon, so that the gun eventually runs out of ammunition and stops firing. While this is happening, the gunner must ensure the weapon stays pointed in a safe direction.

In 2007, a South African anti-aircraft gun malfunctioned on a firing range, resulting in a runaway gun that killed nine soldiers. Contrary to news stories of a “robo-cannon rampage,” the remote gun was not an autonomous weapon and likely malfunctioned because of a mechanical problem, not a software glitch.\(^447\) According to sources knowledgeable about the weapon, it was likely bad luck, not deliberate targeting, that caused the gun to swivel towards friendly lines when it malfunctioned.\(^448\) Unfortunately, despite the heroic efforts of one artillery officer who risked her life to try to stop the runaway gun, the gun poured a string of 35mm rounds into a neighboring gun position, killing the soldiers present.\(^449\)

Runaway guns can be deadly affairs even with simple machine guns that can’t aim themselves. A runaway gun with an autonomous weapon would be a far more dangerous situation. An autonomous weapon that slipped out of control could cause significant destruction.


Imagine a fully autonomous version of the Patriot that operated for an extended period of time without any ability for humans to intervene. “The machine doesn’t know it’s making a mistake,” Hawley observed.450 A single fratricide could become many, with the system continuing to engage inappropriate targets without ceasing. In a runaway machine gun, the gunner can break its links to stop the flow of ammunition. An autonomous weapon could continue engaging targets until it exhausted its ammunition. The consequences to civilians or friendly forces could be disastrous.

The danger of autonomous weapons

By deploying autonomous weapons, militaries are like Mickey enchanting the broomstick. They are trusting that autonomous weapons will perform their functions correctly. They are trusting that they have designed the system, tested it, and trained the operators correctly. They are trusting that the operators are using the system the right way, in an environment they can understand and predict, and that they remain vigilant and don’t cede their judgment to the machine. Normal accident theory would suggest that they should trust a little less.

450 Hawley, interview.
Autonomy is tightly bounded in weapons today. Fire and forget missiles cannot be recalled once launched, but their freedom to search for targets in space and time is limited. This restricts the damage they could cause if they fail. In order for them to strike the wrong target, there would need to be an inappropriate target that met the seeker’s parameters within the seeker’s field of view for the limited time it was active. Such a circumstance is not inconceivable. In fact, that appears to have been what occurred in the F-18 Patriot fratricide. If missiles were made more autonomous, however – if the freedom of the seeker to search in time and space were expanded – the possibility for more accidents like the F-18 shootdown would expand.

Human supervised autonomous weapons such as the Aegis have more freedom to search for targets in time and space, but this freedom is compensated for by the fact that human operators have much more immediate and direct control over the weapon system. Human operators supervise the weapon’s operation in real time. In the Aegis, they can even engage hardware-level cutouts that will disable the power, preventing a missile launch. An Aegis is a dangerous dog kept on a tight leash.

Offensive fully autonomous weapons are different. In the event of a failure, if humans could not reassert control, a fully autonomous weapon’s damage potential would be limited only by its range, endurance, ability to sense targets, and magazine capacity. Loitering munitions are a type of offensive fully autonomous weapon, but their damage potential is relatively limited. While they have a much greater freedom in time and space to search for targets than homing munitions, they are still single-shot weapons that can strike only one target apiece. Autonomous robotic platforms, such as uninhabited combat aircraft or undersea vehicles, could carry more weapons and would have a much higher damage potential. This means the risk of deploying such systems is much greater.

Moreover, militaries rarely deploy single weapons individually. They are likely to deploy large numbers of systems, squadrons of aircraft and fleets of ships. Flaws in any one system will be replicated in all of them, opening the door to what Borrie described as “incidents of mass lethality.” This is fundamentally different from human mistakes, which tend to be idiosyncratic. Hawley said, “If you put someone else in [a fratricide situation], they probably would assess the situation differently and they

451 Borrie, interview.
may or may not do that.” Autonomous systems are different. Not only will it continue making the same mistake, but all other systems of that same type will do so as well.

A frequent refrain in debates about autonomous weapons is that humans also make mistakes and if the machines are better, then we should use the machines. This is a red herring and misunderstands the nature of autonomous weapons. If there are specific engagement-related tasks that automation can do better than humans, then those tasks should be automated. Keeping humans in the loop or at least on the loop provides a vital additional fail-safe, however. It’s the difference between a pilot flying an airplane on autopilot and an airplane with no human in the cockpit at all. The key factor to assess with autonomous weapons isn’t whether the system is better than a human, but rather if the system fails (which it inevitably will), what is the damage it will cause and is that risk acceptable?

Putting an offensive fully autonomous weapon platform into operation would be like turning an Aegis to Auto-Special, rolling FIS green, pointing it towards a communications-denied environment, and having everyone on board exit the ship. Deploying autonomous weapons would be like putting a whole fleet of these systems into operation. The degree of trust that would be required to delegate that amount of lethality to autonomous systems without any ability for humans to intervene is unprecedented.

I asked Captain Galluch what he thought of an Aegis operating on its own with no human supervision. It was the only question I asked him in our four-hour interview for which he did not have an immediate answer. It was clear that in his thirty-year career it had never once occurred to him to turn an Aegis to Auto-Special, roll FIS green, and have everyone onboard exit the ship. He leaned back in his chair and looked out the window. “I don’t have a lot of good answers for that,” he said. But then he began to walk through what one might need to do to build trust in such a system, applying his decades of experience with Aegis. One would need to “build a little, test a little,” he said. High-fidelity computer modeling coupled with real world tests and live fire exercises would be necessary to understand the system’s limitations and the risks of using it. Still, if the military did deploy a fully autonomous weapon, “we’re going to get a Vincennes-like response” in the beginning, he said. The reason is because there is no way to anticipate everything in testing. “Understanding the complexity of Aegis has been a 30-year process,” Galluch said. “Aegis today is not the Aegis of Vincennes,” but

---

452 Hawley, interview.
only because the Navy has learned from mistakes. With a fully autonomous weapon, we’d be starting at year zero.\footnote{Galluch, interview.}

**Nuclear Weapons as a Case Study in Risk Management**

Fully autonomous weapons concentrate destructive capability in the weapon, increasing the consequences if the system were to suffer a failure. In short, full autonomy makes the weapon inherently more dangerous. Are militaries likely to be able to manage these weapons safely? Nuclear weapon safety and accident risk is a valuable case study for evaluating militaries’ ability to safely manage hazardous weapons. If militaries have been able to safely manage nuclear weapons, which are complex and have enormous destructive potential, then they may also be able to safely manage fully autonomous weapons, even if they are susceptible to accidents. However, if states have struggled with nuclear weapons safety and have a history of near-miss accidents (as normal accident theory would predict), then that would lead to a more pessimistic conclusion about militaries’ ability to manage risks from fully autonomous weapons.

Nuclear weapons safety and near-miss accidents have been extensively studied elsewhere, including by Patricia Lewis et al., in “Too Close for Comfort: Cases of Near Nuclear Use and Policy Options” (2014) and Scott Sagan in *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (1993). A brief summary of nuclear near-use incidents and the findings from these studies are included below.

The destructive power of nuclear weapons defies easy comprehension. A single *Ohio*-class ballistic missile submarine can carry 24 Trident II (D5) ballistic missiles, each with eight 100-kiloton warheads per missile. Each 100-kiloton warhead is over six times more powerful than the bomb dropped on Hiroshima. Thus, a single submarine has the power to unleash over 1,000 times the destructive power of the attack on Hiroshima. Individually, nuclear weapons have the potential for mass destruction. Collectively, a nuclear exchange could destroy human civilization. But outside of testing they have not been used, intentionally or accidentally, since 1945.

On closer inspection, however, the safety track record of nuclear weapons is less than inspiring. In addition to the Stanislav Petrov incident in 1983, there have been a number of nuclear near miss incidents that could have had catastrophic consequences.
Some of these could have resulted in an individual weapon’s use, while others could potentially have led to a nuclear exchange between superpowers.

In 1979, a training tape left in a computer at the U.S. military’s North American Aerospace Defense Command (NORAD) led military officers to initially believe that a Soviet attack was underway, until it was refuted by early warning radars. Less than a year later in 1980, a faulty computer chip led to a similar false alarm at NORAD. This incident progressed far enough that U.S. commanders notified National Security Advisor Zbigniew Brzezinski that 2,200 Soviet missiles were inbound to the United States. Brzezinski was about to inform President Jimmy Carter before NORAD realized the alarm was false.

Even after the Cold War ended, the danger from nuclear weapons did not entirely subside. In 1995, Norway launched a rocket carrying a science payload to study the aurora borealis that had a trajectory and radar signature similar to a U.S. Trident II submarine-launched nuclear missile. While a single missile would not have made sense as a first strike, the launch was consistent with a high-altitude nuclear burst to deliver an electromagnetic pulse to blind Russian satellites, a prelude to a massive U.S. first strike. Russian commanders brought President Boris Yeltsin the nuclear briefcase, and Yeltsin discussed a response with senior Russian military commanders before the missile was identified as harmless.

In addition to these incidents are safety lapses that might not have risked nuclear war but are troubling nonetheless. In 2007, for example, a U.S. Air Force B-52 bomber flew from Minot Air Force Base to Barksdale Air Force Base with six nuclear weapons aboard without the pilots or crew being aware. After it landed, the weapons remained onboard the aircraft, unsecured and with ground personnel unaware of the weapons, until they were discovered the following day. This incident was merely the most egregious in a series of recent security lapses in the U.S. nuclear community that caused Air Force leaders to warn of an “erosion” of adherence to appropriate safety standards.

---


Nor were these isolated cases. There were at least 13 near-use nuclear incidents from 1962 to 2002.\(^{458}\) This track record does not inspire confidence. Indeed, they lend credence to the view that near-miss incidents are normal, if terrifying, conditions of nuclear weapons. The fact that none of these incidents led to an actual nuclear detonation, however, presents an interesting puzzle: Do these near-miss incidents support the pessimistic view of normal accident theory that accidents are inevitable? Or does the fact that they didn’t result in an actual nuclear detonation support the more optimistic view that high-reliability organizations can safely operate high-risk systems?

Stanford political scientist Scott Sagan undertook an in-depth evaluation of nuclear weapons safety to answer this very question. He evaluated a number of different incidents to determine whether normal accident theory or high-reliability theory better explained the behavior of those charged with safeguarding nuclear weapons. At the conclusion of his exhaustive study, published in *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*, Sagan wrote:

> When I began this book, the public record on nuclear weapons safety led me to expect that the high reliability school of organization theorists would provide the strongest set of intellectual tools for explaining this apparent success story. … The evidence presented in this book has reluctantly led me to the opposite view: the experience of persistent safety problems in the U.S. nuclear arsenal should serve as a warning.\(^{459}\)

Sagan concluded, “the historical evidence provides much stronger support for the ideas developed by Charles Perrow in *Normal Accidents*” than high-reliability theory. Beneath the surface of what appeared, at first blush, to be a strong safety record was, in fact, a “long series of close calls with U.S. nuclear weapon systems.”\(^{460}\) This is not because the organizations in charge of safeguarding U.S. nuclear weapons were unnaturally incompetent or lax. Rather, the history of nuclear near misses simply

\(^{458}\) Lewis, “Too Close for Comfort.”


\(^{460}\) Ibid, 252.
reflects “the inherent limits of organizational safety.” Military organizations have other operational demands they must accommodate beyond simply safety. Political scientists have termed this the “always/never dilemma.” Militaries of nuclear-armed powers must always be ready to launch nuclear weapons at a moment’s notice and deliver a massive strike against their adversaries, in order for deterrence to be credible. At the same time, they must never allow unauthorized or accidental detonation of a weapon. Sagan says this is effectively “impossible.” There are limits to how safe hazards can be made.

Summary: The Inevitability of Accidents

Safety is challenging enough with nuclear weapons. Autonomous weapons would be potentially more difficult in a number of ways. Nuclear weapons are only available to a handful of actors, but autonomous weapons could proliferate widely, including to countries less concerned about safety. Autonomous weapons have an analogous problem to the always/never dilemma: once put into operation, they are expected to find and destroy enemy targets and not strike friendlies or civilian objects. Unlike nuclear weapons, some isolated mistakes might be tolerated with autonomous weapons, but gross errors would not.

The fact that autonomous weapons are not obviously as dangerous as nuclear weapons might make risk mitigation more challenging in some respects. The perception that automation can increase safety and reliability – which is true in some circumstances – could lead militaries to be less cautious with autonomous weapons than even other conventional weapons. If militaries cannot reliably institute safety procedures to control and account for nuclear weapons, their ability to do so autonomous weapons is far less certain.

The overall track record of nuclear safety, Aegis operations, and the Patriot fratricides suggests that sound procedures can reduce the likelihood of accidents but can never drive them to zero. By embracing the principles of high-reliability organizations, the U.S. Navy submarine and Aegis communities have been able to manage complex,

461 Ibid, 279.
462 Ibid, 278.
463 Ibid, 278.
464 Thanks to Heather Roff for pointing out this parallel.
hazardous systems safely, at least during peacetime. Had the Patriot community adopted some of these principles prior to 2003, the fratricides might have been prevented. At the very least, the Tornado shootdown could have been prevented with a greater cultural vigilance to respond to near-miss incidents and correct known problems, such as the anti-radiation missile misclassification problem, which had come up in testing. High-reliability theory does not promise zero accidents, however. It merely suggests that very low accident rates are possible. Even in industries where safety is paramount, such as nuclear power plants and space travel, accidents still occur.

There are reasons to be skeptical of the ability to achieve high-reliability operations for autonomous weapons, though. High-reliability organizations depend on three key features that work for Aegis in peacetime but are unlikely to be present for fully autonomous weapons in war.

First, high-reliability organizations are able to achieve low accident rates because they are able to constantly refine their operations and learn from near-miss incidents. This is only possible if they can accumulate extensive experience in their operating environment. For example, when Aegis first arrives to an area, the ship operates for a period of time with its radar on and doctrine enabled, but the weapons deactivated, so sailors can see how the doctrine responds to the unique peculiarities of that specific operating environment. Similarly, FAA air traffic control, nuclear power plants, aircraft carriers, and submarines are systems people operate day-in and day-out, accumulating large amounts of operational experience. This daily experience in real-world conditions allows them to refine safe operations.

When there are extreme events outside the norm, safety can be compromised. Users are not able to anticipate all of the possible interactions that may occur under atypical conditions. The 9.0 magnitude earthquake in Japan that led to the Fukushima-Daiichi meltdown is one such example. If 9.0 magnitude earthquakes causing 40-foot high tsunamis were a regular occurrence, nuclear power plant operators would have quickly learned to anticipate the common-mode failure that knocked out primary and backup power. They would have built higher flood walls and elevated the backup diesel generators off the ground. It is difficult, however, to anticipate the specific failures that might occur during atypical events. Accidents such as the 737 MAX crashes are not unsurprising when complex systems are first introduced into real-world environments.

War is an atypical condition. Militaries prepare for war, but the usual day-to-day experience of militaries is peacetime. Militaries attempt to prepare for the rigors of war through training, but no amount of training can replicate the violence and chaos of actual combat. This makes it very difficult for militaries to accurately predict the
behavior of autonomous systems in war. Even for Aegis, activating the doctrine with the weapons disabled allows the operators to understand only how the doctrine will interact with a peacetime operating environment. A wartime operating environment will inevitably be different and raise novel challenges. The USS Vincennes accident highlights this problem. The Vincennes crew faced a set of conditions that were different from peacetime – military and commercial aircraft operating in close proximity from the same airbase coupled with an ongoing hostile engagement from Iranian boats firing at the Vincennes. Had they routinely faced these challenges, they might have been able to come up with protocols to avoid an accident, such as staying off the path of civilian airliners. However, their day-to-day operations did not prepare them – and could not have prepared them – for the complexities that combat would bring. Hawley remarked, “You can go through all of the kinds of training that you think you should do … what nails you is the unexpected and the surprises.”

Another important difference between peacetime high-reliability organizations and war is the presence of adversarial actors. Safe operation of complex systems is difficult because bureaucratic actors have other interests that can sometimes compete with safety – profit, prestige, etc. However, none of the actors are generally hostile to safety. The risk is that people take shortcuts, not actively sabotage safe operations. War is different. War is an inherently adversarial environment in which there are actors attempting to undermine, exploit, or subvert systems. Militaries prepare their troops for this environment not by trying to train their troops for every possible enemy action, but by inculcating a culture of resiliency, decisiveness, and autonomous execution of orders. Humans are expected to adapt in war and come up with novel solutions to respond to enemy actions. This is an area in which humans excel, but machines perform poorly. The brittleness of automation is a major weakness when it comes to responding to adversary innovation. Once an adversary finds a vulnerability in an autonomous system, he or she is free to exploit it until a human realizes the vulnerability and either fixes the system or adapts its use. The system itself cannot adapt. The predictability that a human user finds desirable in automation can be a vulnerability in an adversarial environment.

Finally, the key ingredient in high-reliability organizations that makes them reliable is people, who by definition are not present in the actual execution of operations.

---

465 Hawley, interview.
by a fully autonomous weapon. People are what makes high-reliability organizations reliable. Automation can play a role for “planned actions,” as William Kennedy explained, but humans are required to make the system flexible, so that operations are resilient in the face of atypical events. Humans put slack in a system’s operations, reducing the tight coupling between components and allowing for judgment to play a role in operations. In fully autonomous systems, humans are present during the design and testing of a system and humans put the system into operation, but humans are not present during actual operations. They cannot intervene if something goes wrong. The organization that enables high reliability is not available – the machine is on its own, at least for some period of time. Safety under these conditions requires something more than high-reliability organizations. It requires high-reliability fully autonomous complex machines, and there is no precedent for such systems. This would require a vastly different kind of machine from Aegis, one that was exceptionally predictable to the user but not to the enemy, and with a fault-tolerant design that defaulted to safe operations in the event of failures.

Given the state of technology today, no one knows how to build a complex system that is 100% fail-safe. It is tempting to think that future systems will change this dynamic. The promise of “smarter” machines is seductive: they will be more advanced, more intelligent, and therefore able to account for more variables and avoid failures. To a certain extent, this is true. A more sophisticated early warning system that understood U.S. nuclear doctrine might have been able to apply something similar to Petrov’s judgment, determining that the attack was likely false. A more advanced version of the Patriot might have been able to take into account the IFF problems or electromagnetic interference and withhold firing on potentially ambiguous targets.

It would be a mistake, however, to think that smarter machines could avoid accidents entirely. New features increase complexity, a double-edged sword. More complex machines may be more capable, but harder for users to understand and predict their behavior, particularly in novel situations. For rule-based systems, deciphering the intricate web of relationships between the various rules that govern a system’s behavior and all possible interactions it might have with its environment quickly becomes impossible. Adding more rules can make a system smarter by allowing it to account for

---

466 Kennedy, interview.
more scenarios, but the increased complexity of its internal logic makes it even more opaque to the user.

Implications of risk, predictability, and hazard for stability

The history of experience with complex, autonomous systems in military and non-military settings suggests that accidents are likely to be a normal consequence of operations with fully autonomous weapons. The conditions that enable high-reliability, namely routine operations and the presence of human operators who can adapt to novel situations, would not exist for fully autonomous weapons. Moreover, fully autonomous weapons would run the risk of a “runaway gun” which could continue engaging targets erroneously without stopping – or worse, a systemic failure of multiple autonomous systems at once. The inherent hazard of fully autonomous weapons is high. Should an accident occur, they run the risk of “incidents of mass lethality” in a novel way that semi-autonomous weapons or human-caused accidents do not.

Part III presented two competing hypothesis about ways in which autonomous weapons could affect stability:

H1: Widespread autonomous weapons would **undermine** crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could initiate, escalate, and/or make it more difficult to terminate conflicts.

H2: Widespread autonomous weapons would **improve** crisis stability, escalation control, and war termination by increasing political leaders’ control over their military forces, reducing the risk of accidents or unauthorized escalatory actions.

The conclusion of the examination of risk, predictability, and hazard in autonomous weapons is that militaries are unlikely to be able to operate fully autonomous weapons to the standards of high reliability, and that accidents are likely to be normal features of fully autonomous weapons in realistic military environments. That is not to suggest that accidents would occur all the time. The vast majority of Patriot-aircraft interactions in Operation Iraq Freedom did not result in fratricide. Nevertheless, the risks inherent in the system, even with humans “in the loop,” resulted in an “unacceptable” fratricide rate
(25% of Patriot-related engagements).\textsuperscript{467} The history of complex, autonomous systems in other settings suggests that the Patriot experience is not an aberration, and that the U.S. Navy’s safety record with Aegis operations is peacetime is not likely to be replicated in wartime settings, including with Aegis.

These findings support the first hypothesis (H1). Unintended lethal engagements are likely to be a natural consequence of fully autonomous weapons. Better testing could reduce these risks, but not eliminate them entirely. In crises or conflicts, these unintended lethal engagements could initiate, escalate, and/or make it more difficult to terminate conflicts, undermining crisis stability, escalation control, and war termination.

The remainder of this paper will explore three key future dimensions of autonomous weapons which could conceivably mitigate these risks. These include technical (Part V), operational (Part VI), and legal or policy (Part VII) approaches.

\textsuperscript{467} Hawley, “Looking Back at 20 Years of MANPRINT on Patriot.”
PART V: LEARNING MACHINES

Learning systems would appear to sidestep many of the problems of automation. They don’t rely on rules. Rather, the system is fed data and then learns the correct answer through experience over time. Some of the most innovative advances in AI are in machine learning, such as deep neural networks. Militaries will want to use learning systems to solve difficult problems, and indeed programs such as DARPA’s TRACE already aim to do so. Testing these systems is even more challenging, however. Incomprehensibility is a problem in complex systems, but it is far worse in systems that learn on their own.

Machine Learning and Neural Networks

Because learning machines don’t follow a set of programmed rules, but rather learn from data, they are effectively a “black box” to designers. Computer programmers can look at the network’s output and see whether it is right or wrong but understanding why the network came to a certain conclusion – and more importantly predicting its failures in advance – can be quite challenging. Then Deputy Secretary of Defense Bob Work specifically called out this problem. “How do you do test and evaluation of learning systems?” he asked rhetorically. He didn’t have an answer; it is a difficult problem.

The problem of verifying the behavior of learning systems is starkly illustrated by the vulnerability of the current class of image classifiers to “adversarial images.” Deep neural networks have proven to be an extremely powerful tool for object recognition, outperforming humans in standard benchmark tests. However, researchers have also discovered that, at least with current techniques, they have vulnerabilities that human lack.

Adversarial images are specially created images that exploit deep neural networks’ vulnerabilities to trick them into misidentifying images and at a high degree of confidence. The problem isn’t that the networks get some objects wrong. They


469 Kaiming He et al., “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.”

generally perform object recognition well, and researchers intentionally create these images through evolutionary algorithms.

The problem is that the way in which the deep neural nets get the objects wrong is bizarre and counterintuitive to humans, falsely identifying objects from meaningless static or abstract shapes in ways that humans never would. This makes it difficult for humans to accurately predict under what conditions the neural net might fail. Because the network behaves in a way that seems totally alien, it is very difficult for humans to come up with an accurate mental model of the network’s internal logic that could predict its behavior. Within the black box of the neural net lies a counterintuitive and unexpected form of brittleness, one that is surprising to even the network’s designers.

This is not a weakness of only one specific network or even one application of deep learning. This vulnerability appears to be replicated across all machine learning systems. In fact, one doesn’t even need to know the specific internal structure of the network in order to fool it.\textsuperscript{471}

This vulnerability of deep neural nets to adversarial images is a major problem. In the near-term, it casts doubt on the wisdom of using the current class of visual object recognition AIs for military applications – or for that matter any high-risk applications in adversarial environments. Deliberately feeding a machine false data to manipulate its behavior is known as a spoofing attack, and the current state-of-the-art image classifiers have a known weakness to spoofing attacks that can be exploited by adversaries. Even worse, the adversarial images can be surreptitiously embedded into normal images in a way that is undetectable by humans.\textsuperscript{472} AI researcher Jeff Clune explained this makes it a “hidden exploit” that could allow an adversary to trick the AI system in a way that is invisible to the human. For example, someone could embed an image into the mottled gray of an athletic shirt, tricking an AI security camera into believing the person wearing the shirt was authorized entry, and human security guards wouldn’t even be able to tell a fooling image being used.\textsuperscript{473}

\begin{flushright}
\textsuperscript{471} Szegedy et al., “Intriguing properties of neural networks.”
\end{flushright}

\begin{flushright}
\textsuperscript{472} Szegedy et al., “Intriguing properties of neural networks.”
\end{flushright}

\begin{flushright}
\textsuperscript{473} Jeff Clune, Associate Professor of Computer Science, University of Wyoming, interview by author, telephone, September 28, 2016.
\end{flushright}
Researchers are only beginning to understand why the current class of deep neural networks is susceptible to this type of manipulation. It appears to stem from fundamental properties of their internal structures. Researchers are working on defenses, but for now there is no known effective method of fully inoculating algorithms against these attacks.

In some settings, the consequences of this vulnerability could be severe. Clune gave a hypothetical example of a stock-trading neural net that read the news. News-reading trading bots appear to already be active on the market, evidenced by sharp market moves in response to news events at speeds faster than what is possible by human traders. If these bots used machine learning to understand text—a technique that has been demonstrated and is extremely effective—then they would be vulnerable to this form of hacking. Something as simple as a carefully crafted tweet could fool the bots into believing a terrorist attack was underway. A similar incident already occurred in 2013 when the Associated Press Twitter account was hacked and sent a false tweet reporting explosions at the White House. Stocks rapidly plunged in response. Eventually, the AP confirmed that its account had been hacked and markets recovered, but what makes adversarial attacks so dangerous is that they could be done in a hidden way, without humans even aware that they are occurring.

Researchers are working on making neural networks more robust to adversarial examples, but even if this could be done Clune said “we should definitely assume”

---


478 Domm, “False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; AP Confirms Its Twitter Feed Was Hacked.”

479 “Deep neural networks are easily fooled.”
that the new AI has some other “counterintuitive, weird” vulnerability that we simply haven’t discovered yet.\textsuperscript{480}

In 2017, a group of scientific experts called JASON tasked with studying the implications of AI for the Defense Department came to a similar conclusion. After an exhaustive analysis of the current state of the art in AI, they concluded:

\begin{quote}
the sheer magnitude, millions or billions of parameters (i.e. weights/biases/etc.), which are learned as part of the training of the net … makes it impossible to really understand exactly how the network does what it does. Thus the response of the network to all possible inputs is unknowable.\textsuperscript{481}
\end{quote}

Part of this is due to the early stage or research in neural nets, but part of it is due to the sheer complexity of the deep learning. The JASON group argued that “the very nature of [deep neural networks] may make it intrinsically difficult for them to transition into what is typically recognized as a professionally engineered product.”\textsuperscript{482}

AI researchers are working on ways to build more transparent AI, but Clune isn’t hopeful. “As deep learning gets even more powerful and more impressive and more complicated and as the networks grow in size, there will be more and more and more things we don’t understand. … We have now created artifacts so complicated that we ourselves don’t understand them.” Clune likened his position to an “AI neuroscientist” working to discover how these artificial brains function. It’s possible that AI neuroscience will elucidate these complex machines, but Clune said that current trends point against it: “It’s almost certain that as AI becomes more complicated, we’ll understand it less and less.”\textsuperscript{483} Trends in advanced AI systems, such as those using deep reinforcement learning, would appear to reinforce this point of view.\textsuperscript{484}

Even if it were possible to make simpler, more understandable AI, Clune argued that it probably wouldn’t work as well as AI that is “super complicated and big and weird.” At the end of the day, “people tend to use what works,” even if they don’t

\textsuperscript{480} Clune, interview.


\textsuperscript{482} Ibid, 28.

\textsuperscript{483} Clune, interview.

understand it. “This kind of a race to use the most powerful stuff – if the most powerful stuff is inscrutable and unpredictable and incomprehensible – somebody’s probably going to use it anyway.”

Clune said that this discovery has changed how he views AI and is a “sobering message.” When it comes to lethal applications, Clune warned using deep neural networks for autonomous targeting “could lead to tremendous harm.” An adversary could manipulate the system’s behavior, leading it to attack the wrong targets. “If you’re trying to classify, target, and kill autonomously with no human in the loop, then this sort of adversarial hacking could get fatal and tragic extremely quickly.”

While couched in more analytic language, the JASON group essentially issued the same cautionary warning to DoD:

it is not clear that the existing AI paradigm is immediately amenable to any sort of software engineering validation and verification. This is a serious issue, and is a potential roadblock to DoD’s use of these modern AI systems, especially when considering the liability and accountability of using AI in lethal systems.

Given these glaring vulnerabilities and the lack of any known solution, it would be extremely irresponsible to use deep neural networks, as they exist today, for autonomous targeting. Even without any knowledge about how the neural network was structured, adversaries could generate fooling images to draw the autonomous weapon onto false targets and conceal legitimate ones. Because these images can be hidden, it could do so in a way that is undetectable by humans, until things start blowing up.

Beyond immediate applications, this discovery should make us far more cautious about machine learning in general. Machine learning techniques are powerful tools, but they also have weaknesses. Unfortunately, these weaknesses may not be obvious or intuitive to humans. These vulnerabilities are different and more insidious than those lurking within complex systems like nuclear reactors. The accident at Three Mile Island might not have been predictable ahead of time, but it is at least understandable after the fact. One can lay out the specific sequence of events and understand how one event led to another, and how the combination of highly

485 Clune, interview.

486 JASON, “Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD,” 27.
improbable events led to catastrophe. The vulnerabilities of deep neural networks are different; they are entirely alien to the human mind. One group of researchers described them as “nonintuitive characteristics and intrinsic blind spots, whose structure is connected to the data distribution in a non-obvious way.” In other words: the AIs have weaknesses that we can’t anticipate and we don’t really understand how it happens or why.

This is a problem and it should give one pause when thinking about the potential for machine learning going forward. Artificial intelligence is not like human intelligence. The problem is not just that it isn’t as good as humans at general intelligence or handling ambiguity. That is certainly true, but the differences go much deeper. At the structural level, current AI methods are inspired by the human brain, but they don’t work like the human brain. Those differences at the structural level matter and they can manifest in surprising ways in the AI’s behavior.

**Artificial General Intelligence**

Autonomous weapons pose risks precisely because today’s narrow AI systems fail at tasks that require general intelligence. Machines can beat humans at chess or Go but cannot enter a house and make a pot of coffee. Image recognition neural nets can identify objects but cannot generate a coherent story about what is happening in a scene. Without a human’s ability to understand context, a stock-trading AI system doesn’t understand that it is destroying its own company.

Artificial general intelligence (AGI) is a hypothetical future AI that would exhibit human-level intelligence across a range of cognitive tasks. AGI could be applied to solving humanity’s toughest problems, including those that involve nuance,

---

487 Szegedy et al., “Intriguing Properties of Neural Networks.”


489 For more on how AI’s inhuman cognition will affect strategy, see Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?”; Payne, “AI, warbot”; and Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence*.

490 Machines have been able to caption images with reasonable accuracy, describing in a general sense what the scene depicts. For an overview of AI abilities and limitations in scene interpretation, see JASON, “Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD,” 10.
ambiguity, and uncertainty. An AGI could, like Stanislav Petrov, step back to consider the broader context and apply judgment.

What it would take to build such a machine is a matter of pure speculation, but there is at least one proof that general intelligence is possible: us. Even if recent advances in deep neural networks and machine learning come up short, eventually an improved understanding of the human brain should allow for a detailed reconstruction, neuron by neuron. Brain imaging is improving quickly, and some researchers believe whole brain emulations could be possible with supercomputers as early as the 2040s.

The path to AGI is tremendously uncertain. It may come from whole brain emulations or – what some scientists see as more likely – an improved understanding of the human brain may allow for better neuro-inspired AI. Or AGI may be hopelessly distant. AI experts disagree wildly on when AGI might be created, with estimates ranging from within the next decade to never. A majority of AI experts predict AGI could be possible by 2040 and likely by the end of the century, but no one really knows.

Andrew Herr, who studies emerging technologies for the Pentagon, observed, “When people say a technology is 50 years away, they don’t really believe it’s possible. When they say it’s 20 years away, they believe it’s possible, but they don't know how it will happen.” AGI falls into the latter category. We know general intelligence is possible because humans have it, but we understand so little of our own brains and our own intelligence that it’s hard to know how far away it is.

Methodological approaches for achieving AGI

Some scientists have criticized the current direction of AI research, much of which is focused on deep learning, as fundamentally misguided and unlikely to result in progress towards general intelligence. Cognitive scientist Gary Marcus has emerged as one of the most vocal critiques of deep learning, arguing that a major “reboot” in the

---


494 Andrew Herr, CEO, Helicase, email to the author, October 22, 2016.
field is needed to build “real intelligence.” Marcus advocates for building machines that possess innate knowledge and structures. Marcus identifies at least 11 (and possibly more) “computational primitives,” such as “representations of objects” and “structured, algebraic representations,” that he argues must be hard-coded into machines in order for them to exhibit true intelligent behavior. While Marcus allows for some role for deep learning, he argues that deep learning alone will not build intelligent machines and instead that researchers should focus on “hybrid” systems that include deep learning as well as “classic, ‘symbolic’ AI, sometimes referred to as GOFAI (Good Old-Fashioned AI).” Conversely, other AI researchers have pushed back strongly on Marcus’ views, arguing that no innate knowledge is needed to achieve intelligence. For example, machine learning pioneer and former head of the Association for the Advancement of Artificial Intelligence Tom Dietterich has said:

I think most [machine learning] people believe that methods for incorporating prior knowledge in the form of symbolic rules (or their probabilistic equivalent) are too heavy-handed and, while very useful from an engineering point of view, don't contribute to a plausible theory of general intelligence.

It is worth noting that as a practical matter, most useful machines in real-world situations will be designed using whatever method is most readily available to achieve reliable performance. Today, this usually takes the form of hybrid systems that use deep learning as well as other approaches. There is no dispute among researchers about limitations of deep learning today and its myriad shortcomings. Rather, the dispute is about the most promising approaches for future research.

While such a debate may seem academic, it has real practical impact for AI researchers insofar as beliefs about the value of different methodological approaches

---


499 Tom Dietterich, as quoted in Marcus, “Innateness, AlphaZero, and Artificial Intelligence,” 5.
drive research funding. AI researcher Yoshua Bengio, who along with Yann LeCun and Geoffrey Hinton is considered one of the “godfathers” of deep learning, has pointed out:

There is obviously a strong reaction against deep learning from various camps who probably feel frustrated that this research program has attracted so much attention, funding and industrial success, in comparison to whatever their pet research theme might have been.500

Marcus himself is heavily influenced by his background in cognitive science and human psychology. Marcus has argued that “humans (and other creatures) are born with significant amounts of innate machinery” and therefore machines also must be designed with innate knowledge structures in order to achieve intelligence.501

Marcus’ views are not new. They reflect a school of thought in AI that focuses on symbolic reasoning, or what is sometimes called “Good Old-Fashioned AI,” but has fallen out of favor in recent years because of the progress in deep learning. Marcus himself has been making the argument for nearly twenty years since his 2001 book, The Algebraic Mind: Integrating Connectionism and Cognitive Science.502 His views are less a reaction to the limitations of deep learning than a stubborn attachment to a particular view of how intelligence ought to be in the face of tremendous evidence to the contrary. What is remarkable is how much Marcus’ views have not changed over two decades, despite progress in deep learning. Marcus’ argument for innateness simply does not hold up against the actual evidence of progress in the field of AI research. Deep learning has significant limitations today and it is entirely possible that new directions are needed to achieve broader, general-purpose intelligence.503 Nevertheless,


503 In particular, the problem of “catastrophic forgetting,” in which a neural network forgets a prior task when trained on a new task, is a major limitation of deep learning today which must overcome in order to build general-purpose learning machines that go beyond narrow, task-specific AI systems today. For more on this problem, see “Continual Learning and Catastrophic Forgetting,” in Zhiyuan Chen and Bing Liu, Lifelong Machine Learning (San Rafael, CA: Morgan & Claypool Publishers, 2018), 55. Researchers are making progress on this task. See Shawn Beaulieu et al., “Learning to Continually Learn,” arXiv, February 21, 2020, https://arxiv.org/pdf/2002.09571.pdf; and Khurram Javed and Martha White, “Meta-Learning
progress to-date suggests a steady move away from innateness and hard-wired symbolic reasoning toward more general-purpose algorithms that can learn from experience.

While most practical real-world applications of deep learning use labeled training data, many of the cutting-edge advances in AI have come from deep reinforcement learning, a method in which machines learn from interacting with their environment. Two of the major AI research powerhouses, DeepMind and OpenAI, have both focused a significant amount of their attention on deep reinforcement learning and with impressive results. Not only have they been able to achieve superhuman performance at games like Go, Starcraft, and Dota 2, they have done so using progressively more general-purpose methods. DeepMind’s original version of AlphaGo, which beat 18-time world champion Lee Sedol in 2016, was first trained on a database of 30 million moves of human expert Go players.504 AlphaGo then refined its performance to superhuman levels through additional self-play. An updated version AlphaGoZero, released the following year, reached superhuman performance without any human training data, playing 4.9 million games against itself.505 Perhaps even more impressive, it surpassed AlphaGo’s performance, defeating AlphaGo 100-0. In 2018, DeepMind went a step further, releasing AlphaZero, a more general-purpose game-playing algorithm that could be trained to reach superhuman performance in Go, chess, and shogi (a Japanese strategy game). AlphaZero did this entirely through self-play and without human training data or hand-coded knowledge of the game except for the rules.506 AlphaZero also soundly defeated pre-existing specialized computer programs for each game.

DeepMind and OpenAI have since doubled down on these successes by using deep reinforcement learning to achieve superhuman performance in the computer games Starcraft II, Dota 2, and Quake III Arena in 2019.507 All three games involve multi-


agent learning, in which multiple AI agents are trained simultaneously and must cooperate to accomplish goals. While supervised learning from human playing data was used to help initially train AI agents for Starcraft II, superhuman performance was achieved for Quake III Arena and Dota 2 without any prior human training data. For all games, no hand-coded knowledge of game strategy was used. For Starcraft II and Dota 2, the AI program interfaced with the game through an application programming interface (API). For Quake III Arena, the AI program’s only input was the pixels and game score.

What is most impressive in that in all of these games, the AI programs exhibit intelligent behavior that demonstrates the kind of intuitive understanding of the game that AI researchers seek without any hand-coded knowledge. In order to play these games effectively, the AI program must effectively be able to “understand” relations between different elements of the game, such as that moving an agent in a certain way or interacting with another game element will produce a particular outcome. Yet the AI programs do so without the kind of “computational priors” that Marcus argues are necessary.508 The intelligent behavior the AI system exhibits is an emergent property of the neural network, arising from the learning process and its experience.

In a 2019 blog post called “The Bitter Lesson,” AI researcher Rich Sutton summarized recent AI progress by saying, “The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”509 Sutton argued that in the short-term, AI researchers can often make progress on a problem by applying hand-crafted human knowledge of the problem to design purpose-built algorithms, “but the only thing that

---


508 It is worth noting that for other games researchers have used hybrid models that combine deep learning with other approaches, as Marcus advocates. For example, researchers at MIT and Harvard achieved superhuman performance at the computer role-playing game The Resistance: Avalon using a hybrid architecture incorporating multi-agent reinforcement learning and deductive reasoning. Jack Serrino et al., “Finding Friend and Foe in Multi-Agent Games,” June 5, 2019, https://arxiv.org/pdf/1906.02330.pdf. Such hybrid approaches are likely to continue to bear fruit for achieving intelligent performance on a variety of tasks, but do not take away from the long-term trend towards more generalizable methods combined with massive compute and data.

matters in the long run is the leveraging of computation.” Essentially, simpler methods that apply vast amounts of data and computational power against a problem will outperform those that rely on expert human knowledge.\footnote{510}

Other areas of AI beyond games reinforce this view. In 2019, OpenAI released GPT-2, a large-scale language model that predicts the next word in a string of text. GPT-2 has no innate knowledge of language concepts. Rather, it is a supervised learning model that “was trained simply to predict the next word in 40GB of Internet text.”\footnote{511} In essence, OpenAI downloaded 8 million web pages and threw a tremendous amount of computational power at it, building a giant neural network with 1.5 billion connections. The result was a language model that generates fake text that is relatively believable, and certainly far better than prior language models.\footnote{512} OpenAI chalked up another deep learning success in October 2019 with the release of a robot hand that could manipulate a Rubik’s cube. The hand learned how to move the cube in simulation, with OpenAI using a massive amount of computing power to simulate 13,000 years of experience.\footnote{513} While the hand fell short of human-performance (it dropped the cube 8 out of 10 times on average), it still outperformed prior robot hands that rely on human-crafted rules for movement. Interestingly, to train the robot hand OpenAI re-used the same deep reinforcement learning algorithm it used for to train an AI program to play Dota 2, a step towards more general-purpose learning methods. OpenAI was also able to build a system that was robust to perturbations in the real-world, such as someone poking the cube with a pen or tying two fingers together, by


\footnote{512}You can interact with GPT-2 at “Talk to Transformer,” \url{https://talktotransformer.com/}.

\footnote{513}OpenAI used 64 NVIDIA V100 GPUs plus 920 worker machines with 32 CPU cores each for the Rubik’s cube trainer, training “continuously for several months at this scale while concurrently improving the simulation fidelity, ADR algorithm, tuning hyperparameters, and even changing the network architecture. The cumulative amount of experience over that period used for training on the Rubik’s cube is roughly 13 thousand years, which is on the same order of magnitude as the 40 thousand years used by OpenAI Five.” Ilge Akkaya et al., “Solving Rubik’s Cube With a Robot Hand,” October 17, 2019, \url{https://arxiv.org/pdf/1910.07113.pdf}. 


using a technique called automatic domain randomization (ADR) that randomized conditions in simulation.\footnote{Akkaya et al., “Solving Rubik’s Cube With a Robot Hand,” 24.} This approach may be a step towards more robust AI programs that can function under real-world conditions.

Each of the recent advances in machine learning is very incremental. None represent anything close to human intelligence. Yet they do represent a form of narrow, domain-specific intelligence. If playing chess or Go requires something we call “intelligence” in humans then AlphaZero has it. Progress towards more general-purpose intelligence will undoubtedly require advances in algorithms, but this progress is likely to continue on its current trajectory towards more general-purpose methods that leverage large amounts of compute and data in order to achieve intelligent behavior.

\textit{The control problem}

Regardless of the methodological approach used, there is good reason to believe that AGI would amplify, rather than reduce, today’s problems in controlling autonomous systems. Like Clune’s “weird” deep neural nets, advanced AI is likely to be fundamentally alien.\footnote{Nick Bostrom, “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents,” 2012, \url{http://www.nickbostrom.com/superintelligentwill.pdf}.} Nick Bostrom has argued that biological extraterrestrials would likely have more in common with humans than machine intelligence. Biological aliens (if they exist) would have presumably developed drives and instincts similar to ours through natural selection. They would likely avoid bodily injury, desire reproduction, and seek the alien equivalent of food, water, and shelter. There is no reason to think machine intelligence would necessarily have any of these desires. Bostrom has argued intelligence is “orthogonal” to an entity’s goals, such that “any level of intelligence could in principle be combined with … any final goal.”\footnote{Ibid.} This means a superintelligent AI could have any set of values, from playing the perfect game of chess to making more paperclips.

On one level, the sheer alien-ness of advanced AI makes many of science fiction’s fears seem strangely anthropomorphic. Skynet starts nuclear war because it believes humanity is a threat to its existence, but why should it care about its own existence?

There is no reason to think that a superintelligent AI would inherently be hostile to humans. That doesn’t mean it would value human life, either. AI researcher Eliezer
Yudkowsky has remarked, “The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.”517 A favorite thought experiment in the AI community is a superintelligent “paperclip maximizer.” Bostrom has described the scenario as “a superintelligence whose top goal is the manufacturing of paperclips.” This seemingly innocuous goal could lead to disaster, “with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities.”518

Steve Omohundro has argued that without safeguards, advanced AI would develop “drives” for resource acquisition, self-improvement, self-replication, and self-protection. These would not come from the AI becoming self-aware or “waking up,” but rather be instrumental sub-goals that any sufficiently intelligent system would develop in pursuit of its final goal. Omohundro explains: “All computation and physical action requires the physical resources of space, time, matter, and free energy. Almost any goal can be better accomplished by having more of these resources.” An AI system would seek to acquire more resources to improve the chances of accomplishing its goals, whatever they are. “Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources,” Omohundro wrote. Similarly, self-preservation would be an important interim goal towards pursuing its final goal, even if the AI did not intrinsically care about survival after its final goal was fulfilled. “[Y]ou build a chess playing robot thinking that you can just turn it off should something go wrong. But, to your surprise, you find that it strenuously resists your attempts to turn it off.”519 Omohundro concluded:

Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else’s safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems.520

---


520 Ibid.
If Omohundro is right, then advanced AI is an inherently dangerous technology. Similar to nuclear power, advanced AI that isn’t controlled and managed properly could spark an uncontrollable chain reaction with devastating effects.

In response to this concern, AI researchers have begun thinking about how to build “friendly AI.” A friendly AI would be one whose goals naturally align with human values, so the AI system doesn’t “want” to exterminate humanity to make more paperclips. What goals would make an AI friendly? The answer is not as simple as it first appears. Even something simple like, “Keep humans safe and happy” could lead to unfortunate outcomes. Stuart Armstrong has given an example of a hypothetical AI that achieves this goal by burying humans in lead-lined coffins connected to heroin drips.521

An AI system that understood context and meaning would understand that isn’t what its programmers meant, but that might not matter. Bostrom has argued “its final goal is to make us happy, not to do what the programmers meant when they wrote the code that represents this goal.”522 The problem is that any rule blindly followed to its most extreme can result in perverse outcomes. AI researchers refer to this as the “King Midas problem.” According to Greek myth, King Midas was offered one wish by the gods and he wished that anything he touched turned to gold. But, having not properly specified his wish, Midas was dismayed to find that his food turned to gold so he starved and that his daughter turned to gold when he touched her. Stories of genies granting wishes that have harmful consequences are ancient, and superintelligent AI can be thought of as a type of powerful genie. The challenge in setting goals for a superintelligent AI is therefore crafting a final goal that cannot be mistaken and twisted around in a perverse way.

The problem of “perverse instantiation”523 of final goals is not merely a hypothetical one. It has come up in numerous simple AI systems over the years that have learned clever ways to technically accomplish their goals, but not in the way


523 Bostrom, Superintelligence, Chapter 8.
human designers intended.\textsuperscript{524} A Tetris-playing bot learned to pause the game just before the final brick so that it would never lose.\textsuperscript{525} Another system learned to delete the file containing the correct answer so that it could achieve a perfect score.\textsuperscript{526}

Philosophers and AI researchers have pondered the problem of what goals to give a superintelligent AI that could not lead to perverse instantiation and they have not come to any particularly satisfactory solution. Stuart Russell has argued “a system that is optimizing a function of $n$ variables … will often set the remaining unconstrained variables to extreme values…. This is essentially the old story of the genie in the lamp, or the sorcerer’s apprentice, or King Midas: you get exactly what you ask for, not what you want.”\textsuperscript{527}

These challenges in learning systems suggest that even more advanced forms of AI that didn’t have the kind of limitations of systems today would likely suffer from their own vulnerabilities and control problems. More advanced forms of AI may solve some of today’s problems with autonomous weapons, but could easily introduce their own, thornier problems.\textsuperscript{528}

\textit{Hostile AI}

Furthermore, in any future with more advanced forms of AI, militaries will have to face adversarial uses of that technology. Malicious applications of AI are inevitable. Powerful AI systems with insufficient safeguards could slip out of control and cause havoc, much like the Internet Worm of 1988. Some actors will build harmful AI deliberately. Even if responsible militaries such as the United States’ eschew irresponsible AI applications, the ubiquity of the technology all but assures that other actors – nation-states, criminals, or hackers – will use AI in risky or deliberately harmful ways. The same AI tools being developed to improve cyber defenses, like the Mayhem from the Cyber Grand Challenge, could also be used for offense. Introspective,


\textsuperscript{525} Tom Murphy VII, “The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel . . . after that it gets a little tricky,” \url{https://www.es.cmu.edu/~tom7/mario/mario.pdf}.

\textsuperscript{526} Lehman et al., “The Surprising Creativity of Digital Evolution.”


\textsuperscript{528} For more on AGI and strategy, see Payne, \textit{Strategy, Evolution, and War}, 193-215.
learning and adaptive software could be extremely dangerous without sufficient safeguards. While Brumley was dismissive of the potential for software to become “self-aware,” he agreed it was possible to envision creating something that was “adaptive and unpredictable …. [such that] the inventors wouldn’t even know how it’s going to evolve and it got out of control and could do harm.”

Militaries will need to prepare for this future, but the appropriate response may not be a headlong rush into more autonomy. In a world of intelligent adaptive malware, autonomous weapons are a massive vulnerability, not an advantage. The nature of autonomy means that if an adversary were to hack an autonomous system, the consequences could be much greater than a system that kept humans in the loop. Delegating a task to a machine means giving it power. It entails putting more trust in the machine, trust that may not be warranted if cybersecurity cannot be guaranteed. A single piece of malware could hand control over an entire fleet of robot weapons to the enemy. Former Secretary of the Navy Richard Danzig has compared information technologies to a “Faustian bargain” because of their vulnerabilities to cyber attack: “the capabilities that make these systems attractive make them risky.” He has advocated safeguards such as “placing humans in decision loops, employing analog devices as a check on digital equipment, and providing for non-cyber alternatives if cyber systems are subverted.” Human circuit breakers and hardware-level physical controls will be essential to keeping future weapons under human control. In some cases, Danzig says “abnegation” of some cyber technologies may be the right approach, forgoing their use entirely if the risks outweigh the benefits.

As AI advances, militaries will have to carefully weigh the benefits of greater autonomy against the risks if enemy malware took control. Brumley advocated thinking about the “ecosystem” in

529 Brumley, interview.

530 Countering intelligent, adaptive systems would pose unique challenges. Experience with AI systems in strategy games to-date, such as Go, chess, poker, and real-time computer strategy games, shows that AI systems sometimes engage in counterintuitive and inhuman behavior that is nevertheless effective. For more on the implications of AI on strategy, including countering AI systems, see Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?”; Payne, “AI, warbot”; and Payne, Strategy, Evolution, and War: From Apes to Artificial Intelligence.


532 Ibid, 21.

533 Ibid, 20.
which future malware will operate. The best military ecosystem of autonomous systems would be one that weighed the relative risks of different approaches and retained humans in the right spots to manage those risks.

---

534 Brumley, interview.
PART VI: OPERATIONAL FACTORS: THE MILITARY VALUE AND POTENTIAL CONSEQUENCES OF AUTONOMOUS WEAPONS

Fully autonomous weapons are inherently more hazardous than semi-autonomous weapons because of their greater destructive potential if something were to go awry. Militaries have a vested interest in managing risk and controlling their own weapon systems. Accidents that cause mass civilian casualties, unintended escalation, or fratricide could undermine military objectives. However, it is hard for militaries to know how much testing is “good enough.” There are other factors that weigh against continual testing of autonomous systems – cost, competitive advantages against adversaries, and bureaucratic incentives to deliver products. Rigorous testing can buy down risk but can never eliminate it entirely. This challenge is complicated by the fact that different nations will also have different perceptions of risk. The fact that the Soviet Union developed a “dead-hand” semi-automatic nuclear response system is indicative of this. How much confidence one has in countries’ ability to safely use autonomous weapons is as much a function of organizational, bureaucratic, and social factors as technical ones. It is an open question whether nations can safely manage this technology, but examples from other fields such as nuclear weapons safety are not reassuring.

The best decision-making system would be one that leverages the advantages of both humans and machines. Hybrid human-machine cognitive systems can leverage the precision and reliability of automation, without sacrificing the robustness and flexibility of human intelligence. Maintaining a human “in the loop” and building only semi-autonomous weapons would reduce the risk of accidents. Human involvement is no guarantee against accidents. Humans were “in the loop” in both of the 2003 Patriot fratricides. But human involvement could mitigate the risk of a runaway autonomous weapon that could lead to multiple unintended engagements. In adopting a “human in the loop” policy, though, militaries would be giving up any advantages that accrue from fully autonomous weapons.

Some commentators have suggested that fully autonomous weapons would confer military advantages so decisive that militaries would be forced to build them, regardless of any risks.535 This section will critically examine these claims with the aim

---

of better understanding what the military advantages and disadvantages are for weapons with semi-autonomous, supervised autonomous, and fully autonomous command-and-control paradigms in realistic military environments. This analysis concludes that while the advantages may be overstated in some cases, there will be settings in which fully autonomous weapons would be advantageous.

Having established that militaries would have some incentives to develop fully autonomous weapons despite their risks, this section will then consider in greater detail the specific consequences of accidents with autonomous weapons on stability and pathways to escalation. This analysis concludes that accidents with autonomous weapons could escalate tensions in a crisis. The speed with which events can unfold in physical space is constrained by the physics of moving ships, aircraft, missiles, and other assets through space, regardless of their degree of automation. This will place some bounds on the speed with which events can spiral out of control, giving some opportunity for human intervention and making a "flash war" unlikely, at least among physical objects alone. This dynamic would be very different in cyberspace, where interactions among cyber systems could happen much faster. Even in circumstances where humans have time to react to events and determine how to respond, accidents could exacerbate tensions to a degree that it is politically difficult for leaders to back down from conflict. Or conversely, accidents could create a casus belli for leaders to go to war if they desired.

**Centaur Warfighting: Human-Machine Teaming**

Humans and automation both have advantages, and human-machine teaming which leverages the unique advantages of each is a better approach than using humans or autonomous systems alone. The best weapon systems would be those that optimally use both humans and automation. Understanding how human-machine teaming might work for weapons engagements requires first disaggregating the different roles a human performs today in engagements. In today’s semi-autonomous weapon systems, humans

---

536 Used here, “centaur” human-machine teaming refers to the integration of human and machine decision-making in a joint cognitive system, not necessarily the more intensive fusing of humans and machines through technologies such as brain-computer interfaces. For an exploration of how fusing humans and machines may change warfare, see Christopher Coker, “Still ‘the human thing’? Technology, human agency and the future of war,” *International Relations*, 32 (1), 2018: 23-38, [http://eprints.lse.ac.uk/87629/1/Coker_Human%20Thing.pdf](http://eprints.lse.ac.uk/87629/1/Coker_Human%20Thing.pdf).
perform three kinds of roles when they select targets for engagement. In some cases, humans perform multiple roles simultaneously.

One role is to act as an essential operator of the weapon system. When humans are performing this function, the weapon cannot accurately and effectively complete engagements without the human operator. A human “painting” a target with a laser to direct a laser-guided bomb onto the target is acting as an essential operator. Without the human holding the laser on the target, the bomb won’t hit it.

In the second kind of role, the human acts as a fail-safe. In these situations, the weapon system could function on its own, but the human is in the loop as a backup. The human can intervene to alter or halt the weapon’s operation if it fails or if circumstances change such that the engagement is no longer appropriate.

In the third kind of role, the human acts as a moral agent. In these situations, the human operator is making value-based judgments about whether the use of force is appropriate. For example, the human might decide whether the military necessity of destroying a target outweighs the potential collateral damage.

An anecdote from the U.S. air campaign over Kosovo in 1999 includes an instructive example of all three roles in action simultaneously:

On 17 April 1999, two F-15E Strike Eagles, Callsign CUDA 91 and 92, were tasked to attack an AN/TPS-63 mobile early warning radar located in Serbia. The aircraft carried AGM-130, a standoff weapon that is actually remotely flown by the weapons system officer (WSO) in the F-15E, who uses the infrared sensor in the nose of the weapon to detect the target. CUDA 91, flown by two captains (Phoenix and Spidey) from the 494th Fighter Squadron, launched on coordinates provided by the Air Operations Center. As the weapon approached the suspected target location, the crew had not yet acquired the [enemy radar]. At 12 seconds from impact, the picture became clearer. . . . [The pilots saw the profile outline of what appeared to be a church steeple.] Three seconds [from impact], the WSO makes the call: “I’m ditching in this field” and steers the weapon into an empty field several hundred meters away. . . . Postflight review of the tape revealed no object that could be positively
identified as a radar, but the profile of a Serbian Orthodox church was unmistakable.\textsuperscript{537}

In this example, the pilots were performing all three roles simultaneously. By manually guiding the air-to-ground weapon they were acting as essential operators. Without the guidance of the pilots, the weapon would not have been accurate or effective. They were acting as fail-safes, observing the weapon while it was in flight and making an on-the-spot decision to abort once they realized the circumstances were different from what they had anticipated. They were also acting as moral agents. They assessed the military necessity of the target as not worth the potential collateral damage to what appeared to be a church.

In a different scenario, however, human operators might only perform some of these roles. A GPS-guided bomb, for example, would not need manual guidance while in flight. If such a bomb was network-enabled, giving operators the ability to abort in-flight and the pilots had the ability to observe the target area immediately prior to impact, they still could perform the roles of moral agents and fail-safes, even if they were no longer essential operators once they launched the weapon.

Other types of automation in non-military settings disaggregate these functions in various ways. Commercial airliners today have automation to perform the essential task of flying the aircraft, with human pilots largely in a fail-safe role, able to intervene in the event the automation fails. A person kept on medical life support has machines performing the essential task of keeping him or her alive, but it is humans making the moral judgment whether to continue life support. As automation becomes more advanced across a range of applications, it will become technically possible to remove the human from the role of essential operator in many circumstances. In fact, automating some of the weapon system’s operation could result in far greater accuracy, precision, and reliability than relying on a human operator. Automating the human’s role as moral agent or fail-safe, however, is far harder and would require major leaps forward in AI that do not appear on the horizon.

\textsuperscript{537} Mike Pietrucha, “Why the Next Fighter will be Manned, and the One After That,” \textit{War on the Rocks}, August 5, 2015, \url{http://warontherocks.com/2015/08/why-the-next-fighter-will-be-manned-and-the-one-after-that/}. 
The role of the human as moral agent and fail-safe

The benefit to ‘centaur’ human-machine teaming is that militaries don’t need to give up the benefits of human judgment in order to get the benefits of automation. It is possible to design systems that use automation to perform the essential tasks better than humans, while still retaining humans in the role of moral agents and fail-safes, which humans do better. The U.S. counter-rocket, artillery, and mortar (C-RAM) system is an example of this approach. The C-RAM automates much of the engagement, resulting in more precision and accuracy, but still keeps a human in the loop as a fail-safe.

The C-RAM is designed to protect U.S. bases from rocket, artillery, and mortar attacks. C-RAM uses a network of radars to automatically identify and track incoming rounds. Because the C-RAM is frequently used at bases where there are friendly aircraft in the sky, the system autonomously creates a “Do Not Engage Sector” around friendly aircraft to prevent fratricide.\(^{538}\) The result is a highly automated system that, in theory, is capable of safely and lawfully completing engagements entirely on its own. However, humans are still kept “in the loop” for final verification of each individual target before engagement. One C-RAM operator described the role the automation and human operators play:

> The human operators do not aim or execute any sort of direct control over the firing of the C-RAM system. The role of the human operators is to act as a final fail-safe in the process by verifying that the target is in fact a rocket or mortar, and that there are no friendly aircraft in the engagement zone. A human operator just presses the button that gives the authorization to the weapon to track, target, and destroy the incoming projectile.\(^ {539}\)

C-RAM effectively has a dual-safety mechanism, with both human and automated safeties. The automated safety tracks friendly aircraft in the sky with greater precision and reliability than human operators could. But a human is still retained in the loop to react to unforeseen circumstances. This model also has the virtue of ensuring that


human operators must take a positive action before each engagement, ensuring human responsibility for each shot.

In principle, an approach along the lines of C-RAM’s blended use of automation and human decision-making is optimal. This allows militaries to add automation without giving up the role of the human as moral agent and fail-safe. The human operator may not be able to necessarily prevent all accidents from occurring (after all, humans make mistakes), but the inclusion of a human in the loop allows the operator to adapt to unanticipated situations and undertake corrective action, if required, between engagements. This dramatically reduces the potential for multiple erroneous engagements.

In order for human operators to actually perform the roles of moral agent and fail-safe, they must be trained for and supported by a culture of active participation in the weapon system’s operation. The type of “unwarranted and uncritical trust in automation” that led to the Patriot fratricides would result in a human in the loop in name only. Meaningful human involvement in engagements requires automation designed so that human operators can program their intent into the machine, human-machine interfaces that provide humans the information they need to make informed decisions, training that requires the operators to exercise judgment, and a culture that emphasizes human responsibility. When these best practices are followed, the result can be safe and militarily effective systems like C-RAM, where automation provides valuable advantages but humans remain in control.

This suggests that some of the arguments in favor of autonomous weapons are slightly off-point. Automation has advantages and could make war more ethical and humane, but humans and automation are not mutually exclusive. It is possible to harness the advantages of automation to reduce civilian casualties in war without giving up human control and responsibility.

The limits of centaur warfighting: speed

There are situations in which the idealized centaur model of human-machine teaming breaks down. This can occur when actions are required faster than humans can react or when communications are denied between the human and machine.

Chess is a useful analogy. Centaur human-machine teams generally result in better decision-making in chess, but it is not an optimal model in timed games where

---

540 Hawley, “Not by Widgets Alone.”
the player has a limited amount of time to make a move. When the time to decide is compressed, the human does not add any value compared to the computer alone, and may even be harmful by introducing errors. This is clearly the case today for high-speed chess games where a player has only thirty to sixty seconds to make a move. Over time, as computers advance, this time horizon is likely to expand until humans no longer add any value regardless of how much time is allowed. (This may already be the case today, with recent innovations like DeepMind’s chess-playing AlphaZero, which reached superhuman levels of performance in chess in four hours.542)

There are also military situations where time is so compressed that humans cannot remain in the loop. While humans remain in the loop for C-RAM, this is not feasible in other settings such as defending ships or vehicles from rocket or missile attack, which is why weapons like Aegis have supervised autonomous.543 As future missiles incorporate more intelligent features, including swarming behavior to overwhelm defenses, these defensive human-supervised autonomous weapons are likely to become even more important.

541 Cowen, “What are Humans Still Good for?”


543 For a full list of these systems, see Appendix C.
In circumstances when speed is needed, humans may not be able to be in the loop but they can remain in supervisory control. This is undoubtedly a weaker form of human control. If the weapon fails, there is a greater risk that it could cause harm before the person is able to regain control. Supervisory control also increases the risk that the human off-loads moral responsibility to the machine. Nevertheless, supervisory control at least provides some human oversight of engagements to act as a fail-safe and provide moral judgments, when necessary. The fact that defensive human-supervised autonomous weapons such as Aegis have been in widespread use for decades by numerous countries around the globe suggests that these risks are manageable, provided the weapons are used in limited, defensive settings to protect human-occupied bases and vehicles. In all of these situations, as an additional backup, humans have physical access to the weapon system so that they could disable it at the hardware level, if need be. Supervisory control is much riskier if people do not have physical access to stop the weapon in the event of a runaway gun. Some accidents have occurred with the systems that do exist, but not catastrophes. A world with larger numbers of defensive human-supervised autonomous weapons is likely to look not much different than today.

There will undoubtedly be offensive settings where speed is also valuable. In those cases, however, speed will be valuable in the execution of attacks, not necessarily

---

544 Mary “Missy” Cummings, Professor in Electrical and Computer Engineering, Duke University, interview by author, Durham, NC, June 1, 2016.
the decision to launch them. For example, swarming missiles will need to be delegated the authority to coordinate their behavior and deconflict targets, particularly if the enemy is another swarm. By definition, however, humans will be able to choose the time and place of attack for offensive uses. For some types of targets, it may not be feasible to have humans select every individual enemy object. This will especially be the case if militaries move to swarm warfare, with hundreds or thousands of robots in a swarm. But there are weapon systems today, such as Sensor Fuzed Weapon and the Brimstone missile, where humans choose a specific group of enemy targets and the weapons themselves deconflict which weapon hits which target. Many of the concerns surrounding autonomous weapons – accountability, responsibility, and risk – are minimized if the human is selecting a known group of targets at a specific location and the machine is only carrying out the attack. Provided there are reliable communications back to a human controller, this concept of operations should be feasible. For example, if an enemy swarm were detected, this would allow a human to authorize an attack on the swarm as a whole, without having to specify each individual element, which would be impractical. This would leverage the advantages of automation in speed of action on the battlefield but keep humans in the loop for deciding which targets to attack.

Degraded communications: humans in the loop?

Human supervision is not possible when there are no communications with the weapon. Communications are challenging in some environments, such as underwater. Adversaries will also seek to jam or disrupt communications. This is likely to be of biggest concern forward in the battlespace in areas that are contested by the enemy. Communications in contested areas is not an all-or-nothing proposition. Communications are likely to be degraded, but not necessarily entirely denied. Advanced militaries have jam-resistant communications, although they are limited in bandwidth and range.\(^{545}\) This could allow a human in a nearby vehicle to remotely remain in the loop to authorize engagements. The type of high-bandwidth, high-definition full motion video that is used from drones today would not be possible in contested environments, but some communications are likely possible. This raises a critical question: how much bandwidth is required to keep a human in the loop?

Not much. As one example, the below screen grab from a video of an F-15 strike in Iraq is a mere 12 kilobytes in size.\textsuperscript{546} While grainy, it clearly shows sufficient resolution to make out individual vehicles. A trained operator could discriminate military-specific vehicles, such as a tank or mobile missile launcher, from dual-use vehicles such as buses or trucks.

Targeting image from F-15 strike in Iraq (12 kilobytes in size)

DARPA’s CODE program has precisely this goal, keeping a human in the loop via a thin-line communications link that could transmit 50 kilobits per second, roughly on part with a 56K modem from the 1990s.\textsuperscript{547} This low-bandwidth communications link could transmit one image of this quality every other second. (One kilobyte equals eight kilobits, so a 12 kilobyte image is 96 kilobits.) This would allow a human to view the target and decide whether or not to authorize an engagement.

\textsuperscript{546} U.S. Air Force, VIRIN: 20030411-F-6992H-001 [video], April 11, 2003, \url{https://www.youtube.com/watch?v=IHCL4_v_z3o}.

Denied communications: semi-autonomous weapons for pre-planned targets

This reduced-bandwidth approach would not work in areas where communications are entirely denied. In such environments, semi-autonomous weapons could engage targets that had been pre-authorized by human controllers, as cruise missiles do today. This would generally only be practical for fixed targets however, since mobile targets will move. For mobile targets, the information human operators have on the target’s location would likely be outdated by the time the weapon arrived at its target. In order to destroy a mobile target, the weapon would need to search over a wide area to find the target on its own (i.e., a fully autonomous weapon). One exception would be if enemy targets had unique individually identifiable signatures. For example, if a human had indications that the submarine B-59 was at a particular location in the ocean at a certain time, a human could order a semi-autonomous weapon to search an area of uncertainty around that location for the unique signature of B-59 to sink it. Provided this area of uncertainty was small enough and the confidence in the uniqueness of the target signature was high enough, this risk would be manageable. Issues like accountability and responsibility would also be clear. A human would have decided to attack B-59, not a machine, making it a semi-autonomous weapon. The biggest problem, particularly for undersea weapons, would be the amount of time the weapon was on its own. Cruise missiles can travel for several hours. If the time from the weapon’s launch to when it reached the target was only a few hours, this would not be much different than today. If the time began to stretch into days or weeks, however, then recallability would become a concern.

Denied communications: self-defense for robotic systems

There are circumstances for which militaries might find full autonomy is desirable in communications-denied environments. One reason would be to give uninhabited vehicles the ability to defend themselves if they are attacked. Some robotic systems will be cheap and expendable, but others will be expensive and would not be disposable. Militaries will want to ensure these robotic vehicles are not defenseless. If there is no communications link to a human, any defenses would need to be fully autonomous. Allowing autonomous self-defense incurs some risk of mishaps or behavioral hacking. For example, someone could fire at the robot to get it to return fire and then hide behind human shields to deliberately cause an incident where the robot kills civilians. There would also be some risk of fratricide or unintended escalation in a crisis. Even rules of engagement (ROE) intended purely to be defensive could lead to interactions between opposing systems that results in an exchange of fire.
Suppose, for example, that two countries send robotic systems (air, ground, or sea) into a contested region that is claimed by both nations. Both nations want to conduct patrols within the region to demonstrate their claim and want to challenge the other country’s claim, but without starting a war. Both robots are programmed to only use force in defense. One side programs their robot to take aggressive actions, but only to shoot back if fired upon. The other robot is programmed to fire a warning shot if an adversary gets too close. One side’s robot makes an aggressive maneuver, the other fires a warning shot, and soon both robots are shooting at one another. Perhaps the incident ends there. Perhaps it escalates.

There is no question that delegating self-defense authority would be risky. However, it is hard to imagine that militaries would be willing to put expensive uninhabited systems in harm’s way and leave them defenseless. Provided the defensive action was limited and proportionate, the risks might be manageable.

It seems unlikely that militaries would publicly disclose rules of engagement their robotic systems use for deciding when to shoot. Some degree of transparency between nations could help manage the risks of crisis escalation, though. A “rules of the road” for how robotic systems ought to behave in contested areas might help minimize the risk of accidents and improve stability overall. Some rules, such as, “if you shoot at a robot, expect it to shoot back” are self-reinforcing. Combined with a generally cautious “shoot second” rule requiring robots to withhold fire unless fired upon, such an approach is likely to be stabilizing overall. If militaries could agree on a set of guidelines for how they expect armed robotic systems to interact in settings where there is no human oversight, this would greatly help to manage a problem that is sure to surface as more nations field weaponized robotic systems.

Denied communications: fully autonomous weapons for mobile targets

Another reason why militaries might want fully autonomous weapons is for hunting mobile targets. Many important targets are mobile, such as ships, tanks, air defense radars, and missile launchers. Many of these systems are mobile precisely because it makes it harder to find and destroy them. In an ideal world, a swarm of robots would search for these targets, relay the coordinates and a picture back to a human controller for approval (as CODE intends), then the swarm would attack only human-authorized targets. If a communication link is not available, however, then fully autonomous weapons could be used.
There is no doubt that such weapons would be militarily useful. They would also be risky. In these situations, there would be no ability to recall or abort the weapon if it failed, was hacked, or was manipulated into attacking the wrong target. Unlike a defensive counter-fire response, the weapon’s actions would not be limited and proportionate. It would be going on the attack, searching for targets. Given the risks that such weapons would entail, it is worth asking whether their military value would be worth accepting those risks.

When I asked Captain Galluch from the Aegis training center what he thought of the idea of a fully autonomous weapon, he asked, “What application are we trying to solve?” It’s an important question. For years, militaries have had the ability to build loitering munitions that would search over a wide area and destroy targets on their own. With a few exceptions like the TASM and Harpy, these weapons have not been developed. There are no known examples of them being used in a conflict. Fully autonomous weapons might be useful for hunting mobile targets in communications-denied areas. It’s hard to make the case for them as necessary, however, outside of the narrow case of immediate self-defense for robotic systems.

The main rationale for building fully autonomous weapons seems to be the assumption that others might do so. Even the most strident supporters of military

---

548 Galluch, interview.
robotics have been hesitant about fully autonomous weapons … unless others build them. In a May 2016 event at the Atlantic Council, Bob Work asked:

If our competitors go to Terminators and we are still operating where the machines are helping the humans and it turns out the Terminators are able to make decisions faster, even if they’re bad, how would we respond?549

This is a valid question. The problem is that it can become a self-fulfilling prophecy.550 The fear that others might build fully autonomous weapons could be the very thing that drives militaries to build them.

Deploying fully autonomous weapon would be a weighty risk, but it might be one that militaries decide is worth taking. Doing so would be entering uncharted waters. Experience with existing human-supervised autonomous weapons such as Aegis is useful, but only to a point. Fully autonomous weapons in wartime would face unique conditions that limit the applicability of lessons from high-reliability organizations. The wartime operating environment is different from day-to-day operations, limiting the usefulness of peacetime experience. Hostile actors are actively trying to undermine safe operations. And no humans would be present at the time of operation to intervene or correct problems. Given that accidents are likely to be normal features of autonomous weapons, how might an accident with an autonomous weapon affect a crisis?

Flash War: Autonomy and Escalation

If Stanislav Petrov’s fateful decision had been automated, the consequences could have been disastrous: nuclear war. Nuclear command and control is a niche application, though. One could envision militaries deploying autonomous weapons in a wide variety of contexts but still keeping a human finger on the nuclear trigger.

Non-nuclear applications still hold risks for accidental escalation. Militaries regularly interact in tense situations that have the potential for conflict, even in peacetime. In recent years, the U.S. military has jockeyed for position with Russian warplanes in Syria and the Black Sea, Iranian fast boats in the Straits of Hormuz, and


550 Payne argues this is the case for the militarization of AI overall and that “the scope for a new arms race is clear.” Payne, “AI, warbot.”
Chinese ships and air defenses in the South China Sea. Periods of brinksmanship, where nations flex their militaries to assert dominance but without actually firing weapons, are common in international relations. Sometimes tensions escalate to full blown crises in which war appears imminent, such as the Cuban Missile Crisis. In such situations, even the tiniest incident can be a trigger for war, like a match thrown on a pile of tinder ready to burn. In 1914, a lone gunman assassinated Archduke Ferdinand, sparking a chain of events that led to World War I. Miscalculation and ambiguity are common in these tense situations, and confusion and accidents can generate momentum towards war. The Gulf of Tonkin incident, which led Congress to authorize the war in Vietnam, was later discovered to be partially false; a purported gun battle between U.S. and Vietnamese boats on August 4, 1964 never occurred.551

Robotic systems are already complicating these situations. In 2013, China flew a drone over the Senkaku islands, a contested pile of uninhabited rocks in the East China Sea that both China and Japan claim as their own. In response, Japan scrambled an F-15 fighter jet to intercept the drone. Eventually, the drone turned around and left, but afterward Japan issued news rules of engagement for how it would deal with drone incursions. The rules were more aggressive than those for intercepting manned aircraft, with Japan stating they would shoot down any drone entering their territory that refused to leave. In response, China fired back, stating that any shootdown of their drones would be an “act of war” and that China would “strike back.”552

It might seem reasonable for nations to assert the right to shoot down unauthorized drones entering their territory. Control over one’s territory is the definition of sovereignty. The problem is that many regions around the world are disputed; multiple countries claim sovereignty over them. This is a long-enduring feature of


international relations, one that predates drones and autonomy, but the introduction of robotic weapons complicates how nations deal with these disputes. As drones have proliferated around the globe, they have repeatedly been used to broach other nations’ sovereignty. North Korea has flown drones into South Korea. Hamas and Hezbollah have flown drones into Israel. Pakistan has accused India of flying drones over the Pakistani-controlled parts of Kashmir (a claim India has denied).

When sovereignty is clear, the typical response has been to simply shoot down the offending drone. Pakistan shot down the alleged Indian drone over Kashmir. Israel has shot down drones sent into its airspace. Syria shot down a U.S. drone over its airspace in 2015. A few months later, Turkey shot down a presumed Russian drone that penetrated Turkey from Syria.

These incidents have not led to larger conflagrations, perhaps in part because sovereignty in these incidents was not in dispute. They were clear cases where a drone was sent into another nation’s airspace. Within the realm of international relations, shooting it down was seen as a reasonable response. This same action could be perceived very differently in contested areas, however, such as the Senkaku Islands where both countries assert sovereignty. In such situations, a country whose drone was shot down might feel compelled to escalate in order to back up their territorial claim. Hints of these incidents have already begun, such as the 2016 incident when China seized a U.S. underwater drone in the South China Sea. China quickly returned it after U.S. protests; other incidents might not be resolved so easily. The United States nearly retaliated with a military strike after Iran shot down a U.S. drone in international


airspace in the Persian Gulf in 2019, with President Trump calling off the strike at the last minute.\textsuperscript{559}

All of these complications are manageable if autonomous systems do what humans expect them to do. Robots may raise new challenges in war, but humans can navigate these hurdles, so long as the automation is an accurate reflection of human intent. The danger is if autonomous systems do something they aren’t supposed to – if humans lose control.

That’s already happened with drones. In 2010, a Navy Fire Scout drone wandered 23 miles off course from its Maryland base toward Washington, DC, restricted airspace before it was brought back under control.\textsuperscript{560} In 2017, an Army Shadow drone flew more than 600 miles after operators lost control, before finally crashing in a Colorado forest.\textsuperscript{561} Not all incidents have ended so harmlessly, however.

In 2011, the U.S. lost control of an RQ-170 stealth drone over western Afghanistan.\textsuperscript{562} A few days later, it popped up on Iranian television largely intact and in the hands of the Iranian military. Reports swirled online that Iran had hijacked the drone by jamming its communications link, cutting off contact with its human controllers, and then spoofing its GPS signal to trick it into landing at an Iranian base.\textsuperscript{563} U.S. sources called the hacking claim “complete bullshit.”\textsuperscript{564} (Although U.S. officials did confirm the drone was theirs.)\textsuperscript{565} Either way – whatever the cause of the mishap – the U.S. lost


control of a highly valued stealth drone, which ended up in the hands of another country.

The consequences of a reconnaissance drone wandering off course are some international humiliation and potentially valuable military technology falling into enemy hands. Loss of control with a lethal autonomous weapon, on the other hand, could be another matter. Even a robot programmed only to shoot in self-defense could still end up firing in situations where humans wished it hadn’t. If another nation’s military personnel or civilians were killed, it may not be that easy to de-escalate tensions.

Heather Roff argues there is validity to the concern about a “flash war.” She is less concerned about an “isolated individual platform,” however. One or two autonomous weapons probably wouldn’t make much of a difference to international security. A world with large numbers of them deployed on both sides might be different. Roff said the real concern is “networks of systems” working together in “collaborative autonomy.” If the visions of Bob Work and others come true, militaries will field flotillas of robot ships, wolfpacks of sub-hunting robots undersea, and swarms of aerial drones. In that world, the consequences of a loss of control could be much greater. Roff warned, “if my autonomous agent is patrolling an area, like the border of India and Pakistan, and my adversary is patrolling the same border and we have given certain permissions to escalate in terms of self-defense and those are linked to other systems … that could escalate very quickly.” An accident like the Patriot fratricides could lead to a firestorm of unintended lethality with autonomous systems killing and destroying when they shouldn’t have. If large numbers of civilians or military personnel are killed, the ensuing reaction could plunge nations inexorably towards war.

Bradford Tousley, Director of DARPA’s Tactical Technology Office that builds robotic systems like Sea Hunter, said he’s discussed these concerns with colleagues. He saw automated trading as a “great analogy” for the challenges of automation in military applications. “What are the unexpected side effects of complex systems of machines that we don’t fully understand?” he asked rhetorically. Tousley noted that while circuit breakers were an effective damage control measure in stock markets, “there’s no ‘time out’ in the military.”

---


567 Bradford Tousley, Director, Tactical Technology Office, DARPA, interview by author, Ballston, VA, April 27, 2016.
As interesting as the analogy was, Tousley wasn’t concerned because the speed dimension was vastly different between stock trading and war. “I don’t know that large scale military impacts are in milliseconds,” he said. (A millisecond is a thousand microseconds.) “Even a hypersonic munition that might go 700 miles in 20 minutes – it takes 20 minutes; it doesn’t take 20 milliseconds.”

The sheer physics of moving missiles, aircraft, or ships through physical space imposes time constraints on how quickly events can spiral out of control, in theory giving humans time to adapt and respond.

The exception, Tousley said, was in electronic warfare and cyberspace, where interactions occur at “machine speed.” In this world, “the speed with which a bad event can happen,” he said, “is milliseconds.”

With this arms race in speed comes grave risks. Stock trading is one example of a field in which competitors have succumbed to allure of speed, developing ever-faster algorithms and hardware to shave microseconds off of reaction times. In uncontrolled, real-world environments, however, the (unsurprising) result has been accidents. When these accidents occur, machine speed become a major liability. Autonomous processes can rapidly spiral out of control, destroying companies and crashing markets. It’s one thing to say that humans will have the ability to intervene, but in some settings, their intervention may be too late. Automated stock trading foreshadows the risks of a world where nations have developed and deployed autonomous weapons.

A flash physical war in the sense of a war that spirals out of control in mere seconds seems unlikely. Missiles take time to move through the air. Sub-hunting undersea robots can only move so quickly through the water. Accidents with autonomous weapons could undermine stability and escalate crises unintentionally, but these incidents would likely take place over minutes and hours, not microseconds. This is not to say that autonomous weapons do not pose serious risks to stability; they do. A runaway gun could push nations closer to the brink of war. If an autonomous weapon (or a group of them) caused a significant number of deaths, tensions could boil over to the point where de-escalating is no longer possible. The speed at which events would unfold, however, is likely one that would allow humans to see what was happening and, at the very least, take steps to attempt to mitigate the effects. Bob Work said he saw a role for a human “circuit breaker” in managing swarms of robotic systems. If the swarm

---

568 Ibid.
569 Ibid.
began to behave in an unexpected way, “they would just shut it down,” he said. There are problems with this approach. The autonomous system might not respond to commands to shut it down, either because it is out of communications or because the type of failure it is experiencing prevents it from accepting a command to shut down. Unless human operators have physical access, like the physical circuit breaker in Aegis, any software-based “kill switch” is susceptible to the same risks as other software – bugs, hacking, or unexpected interactions.

Even though accidents with physical autonomous weapons will not cascade into all-out war in mere seconds, machines could quickly cause damage that might have irreversible consequences. Countries may not believe that an enemy’s attack was an accident, or the harm may be so severe that they simply don’t care. If Japan had claimed that the attack on Pearl Harbor was not authorized by Tokyo and was the work of a single rogue admiral, it’s hard to imagine the United States would have refrained from war.

A flash cyber war, on the other hand, is a real possibility. Automated hacking back could lead to escalation between nations in the blink of an eye. In this environment, human oversight would be merely the illusion of safety. Automatic circuit breakers are used to stop flash crashes on Wall Street because humans cannot possibly intervene in time. There is no equivalent referee to call time out in war.

**Arms Races and the Security Dilemma**

In 2015, a group of prominent AI and robotics researchers signed an open letter calling for a ban on autonomous weapons. “The key question for humanity today,” they wrote, “is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable.”

The traditional understanding of an arms race does not accurately describe the situation with autonomous weapons today, however. Michael D. Wallace, in his 1979 article “Arms Races and Escalation,” defined an arms race as “involving simultaneous abnormal rates of growth in the military outlays of two or more nations” resulting from “the competitive pressure of the military itself, and not from domestic forces exogenous

---

570 Work, interview.

to this rivalry.” Wallace further stated that the concept of an arms race only applied “between nations whose foreign and defense policies are heavily interdependent” and who have “roughly comparable” capabilities.\textsuperscript{572} Wallace distinguished arms races from the normal behavior of states to improve their military forces. The decisive factor, in Wallace’s determination, was the rate of growth in defense spending. Wallace characterized arms races as resulting in abnormally large growth rates in defense spending, beyond the historical average of 4% to 5% growth (in real dollars), with arms races resulting in annual growth rates above 10% or even as high as 20-25%.\textsuperscript{573} Other scholars define arms races using different quantitative thresholds – and some definitions lack clear quantitative thresholds at all – but the existence of rapid increases in defense spending or military forces above normal levels is a common criteria.\textsuperscript{574}

Arms races result in situations in which two or more countries are locked in spiraling defense spending, grabbing ever-greater shares of national treasure often with little to no net gain in relative advantage over the other. Yet this is clearly not the case today with autonomous weapons. Instead, state behavior with regard to autonomous weapons looks much more like hedging. Militaries are investing in robotics, autonomy, and AI, which has many uses. Each generation of weapons incorporate greater autonomy. But states are not rushing to develop autonomous weapons. Spending on AI and robotics among most militaries is a fraction of defense spending on traditional human-inhabited assets. Most militaries could accelerate robotics development if they chose. Both France and Great Britain and have been taking their time operationalizing UCAV prototypes that flew in 2012 and 2013, respectively. Both nations now envision operational UCAVs as a complement to a new manned fighter aircraft in the 2030s to 2040s.\textsuperscript{575} The United States, similarly, cancelled its plans for a carrier-based UCAV.

\textsuperscript{572} Michael D. Wallace, “Arms Races and Escalation,” \textit{Journal of Conflict Resolution} 23 (1), March 1979, 5.

\textsuperscript{573} Wallace, “Arms Races and Escalation,” 6.


Most major military powers haven’t foresworn autonomous weapons. They are keeping their options open. China’s carefully crafted position at the UN that allows China to claim it supports a ban while effectively restricting nothing is a case in point. Of the nearly 30 nations who have publicly said they support a ban, none of them are leading robotics developers or major military powers. States who support a ban aren’t giving up anything as a practical matter, at least not in the near term.

Nor do autonomous weapons seem like the kind of weapon that would be likely to manifest in a classical arms race that leads to spiraling defense expenditures, simply due to the fact that they are not particularly expensive. Quite the opposite, in fact. Once the autonomy is developed, it can be replicated fairly cheaply. While there may be significant up-front costs in software development, converting a semi-autonomous weapon to a fully autonomous weapon could be done via a software patch, meaning that it could scale very inexpensively. If militaries decided to build massive fleets of tens of thousands of robotic combat vehicles, producing the physical vehicles would be costly but switching from human-in-the-loop (semi-autonomous) to human-out-of-the-loop (fully autonomous) would not be.

Instead, the situation states find themselves in is much more accurately described as a security dilemma, a more generalized competitive dynamic between states than the more narrowly defined arms race. In his 1978 article, “Cooperation Under the Security Dilemma,” Robert Jervis defined the security dilemma as the problem that “many of the means by which a state tries to increase its security decrease the security of others.”^576 As Charles Glaser has pointed out, it is not obvious from this definition why this would be intrinsically a bad thing for states, that an increase in their security comes at the expense of others.\(^577\) The problem is the second- and third-order effects that could develop from another state’s reaction to having their security reduced. While it is possible that the net effect is simply counter-balancing and no change in security, or possibly still an increase in a state’s security if rival state cannot innovate or modernize their military forces as quickly, Glaser also outlines a number of circumstances that could lead to a net negative outcome that is worse off for all states.


One of the ways that the security dilemma could leave states worse off than before is if military innovation and the development of new capabilities pushes warfare into a domain that is more harmful, destructive, less stable, or otherwise less desirable than before. In his 1997 article, “The Security Dilemma Revisited,” Glaser gave an example of capabilities that shifted warfare to a more offense-dominant regime. This could be one way in which warfare could evolve in a net negative direction, but there are others as well. In World War I, Germany was spurred to develop and deploy chemical weapons in part due to fears about French developments in poison gas. The result was the introduction of a weapon that increased combatant suffering on all sides, without delivering a significant military advantage for either side. Autonomous weapons have the potential to do something similar, not by shifting the offense-defense balance but by moving warfare to a new domain with less human control.

All states would be better off if there were no autonomous weapons. There are advantages to adding intelligence into machines, but the optimal model for highest quality decision-making would be a centaur approach that combined human and machine decision-making in weapon systems. Yet it is the fear of falling behind other nations, more than anything else, that drives nations to invest in greater autonomy and, in particular, greater speed. This is a classic security dilemma. One state’s pursuit of greater autonomy and, in particular, faster reaction times undermines other states’ security and leads others to similarly pursue more autonomy just to keep up. Then Deputy Secretary of Defense Robert O. Work summed up the dilemma when he stated, “If our competitors go to Terminators and we are still operating where the machines are helping the humans and it turns out the Terminators are able to make decisions faster, even if they’re bad, how would we respond?” Then Vice Chairman of the Joint Chiefs of Staff General Paul Selva reportedly termed this problem the “Terminator conundrum.” It could be the fear of falling behind others, more than anything else, that motivates states to pursue autonomous weapons.

---


580 Work, remarks at the Atlantic Council Global Strategy Forum.

If states fall victim to this trap, the net result is likely to be an outcome that leaves all states less secure, since the pursuit of greater autonomy would not merely be an evolution in weapons and countermeasures that simply lead to different future weapons. An autonomous weapon is not like a longer spear or a new tank with thicker armor. At some point, warfare shifts into a qualitatively different space where humans have less control over lethal force, as decisions become more automated and the accelerating tempo of operations pushes humans out of the loop. Some Chinese scholars have hypothesized about a “battlefield singularity” in which the pace of combat action eclipses human decision-making. Some U.S. scholars have used the term “hyperwar” to refer to a similar situation. While the speed of engagements necessitates automation in some limited areas today, such as immediate localized defense of ships, bases, and vehicles from rocket and missile attack, an expansion of this zone of machine control into broader areas of war would be a significant development. An era where humans have less control over warfare could lead to wars that are less controllable and escalate more quickly or more widely than humans intend. The net effect of the natural and quite rational desire for nations to pursue an edge in speed could lead to an outcome that is worse for all. And yet competitive dynamics could nevertheless drive such a result.

This dynamic of a competition in speed is like an arms race, if we expand the definition of an arms race to be more in line with biological examples of competitive co-evolution. Biologists often use the metaphor of an arms race to explain “an unstable runaway escalation” of adaptation and counter-adaption that can occur in animals. This can occur between species, such as predator and prey, or within species, such as males evolving in competition for females. Biological arms races can manifest in a variety of ways, such as competitions between predator and prey for greater speed, camouflage vs. detection, armor vs. claws, cognitive abilities, poison, deception, or


other attributes that might advantage survival.\(^{585}\) This broader biological definition of an arms race is more in line with the potential for an escalating “arms race in speed” among nations that leads to greater autonomy in weapons. While this concept does not meet the traditional definition of an arms race in the security studies literature, it is nevertheless a useful concept to describe the potential for a co-evolution in autonomy and speed that leads to no net relative advantage and in fact leaves both sides worse off.

How should nations respond to this security dilemma? Unlike the evolution of animal species, states have the ability to reflect on their situation and make choices about their actions. Michael Carl Haas, who has raised concerns about crisis stability, suggested that countries explore “mutual restraint” as an option to avoid the potentially dangerous consequences of fully autonomous weapons.\(^{586}\) Others have suggested that autonomous weapons are inevitable.\(^{587}\) The history of arms control provides valuable insights on what forms of restraint might be feasible with autonomous weapons.

---

\(^{585}\) Dawkins and Krebs, “Arms races between and within species,” 489-498.


\(^{587}\) Ackerman, “We Should Not Ban ‘Killer Robots.’

264
PART VII: RISK MITIGATION: POLICY OPTIONS AND HISTORICAL PRECEDENTS

There have been many attempts to control weapons in the past. Some have succeeded, but many attempts at restricting weapons have failed. These successes and failures provide lessons for those who wish to control autonomous weapons. There are over 40 cases of historical attempts to ban weapons (Appendix D) with a mixed track record of success. An analysis of why bans succeed or fail identifies three key factors that determine the feasibility of mutual restraint, as well as some secondary factors. This section then compares four distinct policy options for regulating autonomous weapons against the conditions that affect regulatory success.

The conclusion of this analysis is that preventing the proliferation of the underlying technology that enables autonomous weapons is not feasible. The technology is too diffuse, commercially-available, and easy to replicate. Mutual restraint among nations on how they implement this technology may be possible, but there are significant hurdles to overcome. There are precedents for nations refraining from weaponizing widely available technologies. The decisive factor will be how valuable autonomous weapons prove on the battlefield. The widespread availability of AI means that some actors will undoubtedly build weaponize it. The question will be whether countries who refrain from certain kinds of autonomous weapons can still defeat those on the battlefield who show no such restraint.

This section concludes that a comprehensive ban on autonomous weapons is unlikely to succeed and presents three alternative policy options that have a greater likelihood of success, although none are assured. This section then concludes with some recommendations for policymakers for reframing issues surrounding autonomous weapons.

Why Some Bans Work and Others Fail

Banning a weapon isn’t easy but isn’t impossible either. Even in the midst of total war, states have often sought mutual restraint on certain weapons or tactics that would take war to a new level of horror. Whether or not a ban succeeds in restraining use in wartime depends principally on three factors: the perceived horribleness of the weapon; its perceived military utility; and the number of actors who need to cooperate
for a ban to work. If a weapon is seen as horrific and only marginally useful, then a ban is likely to succeed. If a weapon brings decisive advantages on the battlefield, then a ban is unlikely to work no matter how terrible it may seem. The difference between how states have treated chemical weapons and nuclear weapons illustrate this point. Nuclear weapons are unquestionably more harmful than chemical weapons by any measure: civilian casualties, combatant suffering, and environmental damage. Nuclear weapons give a decisive advantage on the battlefield, though, which is why the Nuclear Non-Proliferation Treaty’s (NPT) goals of global nuclear disarmament remain unrealized.

The result of this dynamic is that many successful bans are against weapons that are not particularly effective militarily. It is overly simplistic to say that if a weapon has value, then a ban is doomed to fail, however. If the only factor that mattered was the battlefield utility of a weapon, then militaries would almost certainly use poison gas. It has value in disrupting enemy operations and is certainly an effective psychological weapon. Even if the only effect of using gas were to force the enemy to fight in protective gear, this alone is valuable, as it slows enemy troops down and reduces their effectiveness. Fighting in a gas mask is hard. It reduces situational awareness, makes it difficult to breathe, and diminishes firing accuracy. Expanding bullets and blinding lasers also have some military utility, yet bans on these weapons have been largely successful. In these cases, the perceived value is low enough that it has not been worth it to states to break these prohibitions.

The number of countries that need to participate in order 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 20th century when all great powers needed to agree. A single defector could cause an arms control agreement to unravel, as the Washington Naval Treaty did. Since the end of the Cold War, this


dynamic has begun to reemerge with the collapse of the Anti-Ballistic Missile (ABM) Treaty and Intermediate-Range Nuclear Forces (INF) Treaty.

There are other secondary factors that also influence the likelihood of a ban’s success. A clear focal point is generally needed for effective coordination. Treaties that completely ban a weapon tend to be more successful than complicated rules governing a weapon’s use. In fact, bans that permit weapons to be used in some circumstances in war but not others have not generally been successful. Bans that attempt to regulate how indiscriminate weapons are used on the battlefield in order to avoid civilian casualties (aerial bombardment, submarine warfare, incendiary weapons, and the

591 Treaty Between The United States of America and The Union of Soviet Socialist Republics on The Limitation of Anti-Ballistic Missile Systems (ABM Treaty).


Convention on Certain Conventional Weapons (CCW) land mine protocol)\(^596\) have had a poor track record of success. Complete bans on weapons (exploding bullets, expanding bullets, chemical weapons, biological weapons, environmental modification weapons, blinding lasers)\(^602\) have fared better.\(^603\)

There are two notable classes of bans that seem to be exceptions to this general rule about the importance of simplicity. The first is the bans on land mines\(^604\) and cluster databases.icrc.org/applic/ihl/ihl.nsf/Article.xsp?action=openDocument&documentId=14FEADA9AF35FA9C12563CD0051EF1E.


\(^599\) Chemical Weapons Convention.


\(^601\) Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques.


\(^603\) This seems to suggest that if lasers were used in future wars for non-blinding purposes and ended up causing incidental blinding, then they would quickly evolve into use for intentional blinding. Chemical Weapons Convention, Organisation for the Prohibition of Chemical Weapons, https://www.opcw.org/chemical-weapons-convention/, accessed June 19, 2017.

munitions. The fact that they are written to have simple policies (“never under any circumstances to use …”) and shift the complicated rules to the definitions suggests that the drafters of these treaties are aware that simplicity aids the normative power of a ban. If the goal is to stigmatize a weapon, that only works if the weapon is seen as illegitimate in all circumstances. The other exception is a series of bilateral arms control treaties between the United States and the Soviet Union / Russia: the INF Treaty, ABM Treaty, Strategic Arms Limitation Talks (SALT) I, SALT II, Strategic Offensive Reductions Treaty (SORT), Strategic Arms Reduction Treaty (START), and New START. These treaties had very specific rules on what was permissible, similar to the 1922 Washington Naval Treaty. The fact that only two parties need to agree on these rules has likely been a factor in their success.

Another factor that may have enabled these more complicated bans to succeed is that they apply to the production, stockpiling, or deployment of weapons, not to use in war. Weapons bans can apply at many different stages in the development and use cycle of a weapon. Non-proliferation regimes aim to prevent access to the underlying technology behind certain weapons. Bans like those on land mines and cluster munitions allow access to the technology, but prohibit developing, producing, or stockpiling the weapon. Other bans only apply to use, sometimes prohibiting use entirely and sometimes only certain kinds of uses. Finally, arms limitations treaties


606 Ibid.


permit use; they simply limit the quantities of certain weapons that countries can have in peacetime. In general, more complicated forms of restraint are possible in peacetime but are less viable in wartime when the pressures to breach the prohibition may be higher.

There is a certain path-dependence to bans and why some weapons are perceived as more terrible than others. The campaign to ban cluster munitions likely succeeded riding on the coattails of the campaign to ban land mines. If the order had been reversed, the outcome for both treaties might have been different. This path dependence can lead to strange outcomes. Expanding bullets are perceived as illegitimate for use in war today because of an 1899 treaty, never mind the fact that they are regularly used for law enforcement and personal self-defense. The original 1899 ban on expanding bullets piggybacked on the 1868 ban on exploding bullets. Riot control agents such as tear gas are banned for use in war because they are captured under the Chemical Weapons Convention,612 not because they are seen as causing unnecessary suffering. The bizarre result is that militaries are permitted to use riot control agents against rioting civilians, but not enemy combatants.613

---


One lesson from history is that it is very hard to predict the future path of technology. The 1899 Hague declarations on balloon-delivered weapons, expanding bullets, and gas-filled projectiles made assumptions about the future path of these technologies that turned out to be wrong. The rules banned some weapons that turned out not to be harmful (expanding bullets) and failed to ban those that were (poison gas in canisters).\textsuperscript{614} Anticipating countermeasures to a weapon, which will develop over time, is also difficult. Turn of the century attempts to regulate air-delivered weapons correctly anticipated the advent of the airplane but not the futility in defending against air attack, since regulations only banned attacks on “undefended” cities.\textsuperscript{615}

Technology will evolve in unforeseen ways. This is a problem for preemptive bans on emerging technologies. The cluster munitions and land mines bans can succeed with complicated rules governing what is and is not permissible because the weapons already exist. History suggests this is a losing strategy for preemptive bans. Successful preemptive bans focus on the intent behind a technology, rather than specific restrictions. For example, the blinding laser ban prohibits lasers specifically designed to cause permanent blindness, rather than limit a certain power level in lasers. The United States takes a similar intent-based interpretation of the expanding bullets ban, that they are prohibited only to the extent that they are intended to cause unnecessary suffering.

Preemptive bans pose unique challenges and opportunities. Because they are not yet in states’ inventories, the military utility of a new weapon, such as blinding lasers or environmental modification, may be amorphous. This can sometimes make it easier for a ban to succeed. States may not be willing to run the risk of sparking an arms race if the military utility of a new weapon seems uncertain. On the other hand, states often may not fully understand how terrible a weapon is until they see it on the battlefield. States correctly anticipated the harm that air-delivered weapons could cause to unprotected cities, but poison gas and nuclear weapons shocked the conscience in ways that contemporaries were not prepared for.

The legal status of a treaty seems to have little to no bearing on its success. Legally-binding treaties have been routinely violated and restraint has existed in some


cases without any formal agreements. International agreements, legally binding or not, primarily serve as a focal point for coordination. What actually deters countries from violating bans is not a treaty, since by default there are no legal consequences if one wins the war, but rather reciprocity. Countries show restraint when they fear that another country might retaliate in kind. When they are fighting nations who do not have the ability to retaliate, they show less restraint.

In peacetime, verification can be an essential tool for ensuring that other nations are complying with their end of the bargain. The NPT, Chemical Weapons Convention, INF Treaty, START, and New START all have procedures for inspections to verify compliance. The Outer Space Treaty has a de facto inspection regime, requiring states to permit others to view space launches and visit any facilities on the Moon. The land mine and cluster munitions bans do not have inspection regimes but do require transparency from states on their stockpile elimination.

Not all successful bans include verification, however. The 1899 ban on expanding bullets, 1925 Geneva Gas Protocol, Convention on Certain Conventional Weapons (CCW), and SORT all do not have verification regimes. There is no also verification regime for the Outer Space Treaty’s ban on putting weapons of mass destruction in orbit. The Environmental Modification Convention and Biological Weapons Convention only say that states concerned that another is cheating should lodge a complaint with the UN Security Council.

Some treaties do not include inspections, but specifically state that nations will verify each other’s compliance through their own observations. The Seabed Treaty permits states to observe the activities of others on the seabed and if a state is concerned


about another’s activities, then refer the matter to the UN Security Council. The SALT I, SALT II, and ABM treaties stated that the U.S. and U.S.S.R. will use their own means of verifying compliance. For these treaties, verification could be done via satellite imagery. The Washington Naval Treaty, which was successful for 14 years, similarly does not have a verification provision. Presumably it would be hard to hide a new fleet of battleships.

In general, verification regimes are useful if there is a reason to believe that countries might be developing the prohibited weapon in secret. That could be the case if they already have it (chemical weapons, land mines, or cluster munitions) or if they might be close (nuclear weapons). Inspection regimes are not always essential. What is required is transparency. Countries need to know whether or not other nations are complying. It was, in part, the fear of French poison gas that led Germany to develop and use it in World War I.\textsuperscript{619} In some cases, the need for transparency can be met by the simple fact that some weapons are difficult to keep secret. Anti-ballistic missile facilities and ships cannot be easily hidden. Other weapons can be.

Finally, the motivation behind a ban seems to matter in terms of the likelihood of success. Successful bans fall into a few categories. The first is weapons that are perceived to cause unnecessary suffering. By definition, these are weapons that harm combatants excessive to their military value. Restraint with these weapons is thus easily mutually self-reinforcing. Combatants have little incentive to use these weapons and strong incentives not to, since the enemy would almost certainly retaliate. Germany unilaterally removed the sawback bayonet, which had a serrated edge for sawing wood, during World War I because of concerns about the severity of wounds it caused. Reportedly, British and French troops were so appalled by the weapon that they would torture and kill any captured German soldier found with one.\textsuperscript{620}

Bans on weapons that were seen as causing excessive civilian harm have also succeeded, but only when those bans prohibit possessing the weapon at all (cluster munitions and the Ottawa land mine ban), not when they permit use in some circumstances (air-delivered weapons, submarine warfare, incendiary weapons, and the CCW land mine protocol). Bans on weapons that are seen as destabilizing (Seabed Treaty, Outer Space Treaty, ABM Treaty, INF Treaty) have generally succeeded, at


At least when only a few parties are needed for cooperation. (Both the ABM Treaty and INF Treaty have collapsed in the post-Cold War multi-polar world.) Arms limitation has been exceptionally difficult, even when there are only a few parties, but has some record of success. Prohibiting the expansion of war into new geographic areas has only worked when the focal point for cooperation is clear and there is low military utility in doing so, such as banning weapons on the Moon or Antarctica. Attempts to regulate or restrict warfare from undersea or the air failed, most likely because the regulations were too nuanced. “No submarines” or “no aircraft” would have been clearer, for example.

Ultimately even in the best of cases, bans aren’t perfect. Even for highly successful bans, there will be some nations who don’t comply. This makes military utility a decisive factor. Nations want to know they aren’t giving up a war-winning weapon. This is a profound challenge for those seeking a ban on autonomous weapons.

**Policy Options and Prospects for Regulation**

Since 2014, countries have met annually at the UN CCW in Geneva to discuss autonomous weapons. This section will analyze the current state of progress in international discussions, then consider four policy options for regulation. The first policy option to be analyzed will be a comprehensive, legally-binding treaty banning autonomous weapons. This is the position of the Campaign to Stop Killer Robots, a consortium of over 60 non-governmental organizations, and has been endorsed by thousands of AI scientists and twenty-seven states (as of November 2018). Three other alternative regulatory frameworks are also examined. All of these will be examined within the context of some of the most critical conditions necessary to achieve successful regulation. Successful examples have a few common characteristics. First, a clear focal point for coordination is needed. Countries will need to know what is in and what is out. The simpler and clearer the line, the better. This means that some rules like “no general intelligence” are dead on arrival. An open letter signed by


622 Campaign to Stop Killer Robots, “Country Views on Killer Robots.”
3,000 AI scientists called for a ban on “offensive autonomous weapons beyond meaningful human control.” Every single one of those words is a morass of ambiguity. If states could agree on the difference between “offensive” and “defensive” weapons, they would have banned offensive weapons long ago. “Meaningful human control” is even more vague. Preemptive bans that try to specify the exact shape of the technology don’t work either. The most effective preemptive bans focus on the key prohibited concept, like banning lasers intended to cause permanent blindness.

Second, the horribleness of a weapon must outweigh its military utility for a ban to succeed. Regardless of whether the weapon is seen as destabilizing, a danger to civilians, or causing unnecessary suffering, it must be perceived as bad enough – or sufficiently militarily useless – that states are not tempted to breach the ban.

Third, transparency is essential. States must trust that others are not secretly developing the weapon they themselves have foresworn. Bob Work said he thought countries “will move toward some type of broad international discussion on how far we should go on autonomous weapons.” The problem he saw was verification: “The verification of this regime is going to be very, very difficult because it’s just – it’s ubiquitous. It’s now exploding around us.” This is a fundamental problem for autonomous weapons. The essence of autonomy is software, not hardware, making transparency very difficult.

There are many different ways countries could restrain themselves. States could agree to refrain from certain types of machine intelligence, such as “online” machine learning that adapts on the battlefield. States could focus on physical characteristics of autonomous weapons: size, range, payload, etc. Examining all of the possible forms of regulation is beyond the scope of this thesis. To illustrate the range of possibilities, below are four very different ways that states could approach this problem. These solutions are not exhaustive of all possibilities, nor are they mutually exclusive.

Option 1: Ban fully autonomous weapons

The Campaign to Stop Killer Robots has called for “a comprehensive, preemptive prohibition on the development, production and use of fully autonomous weapons.” Assuming that states found it in their interests to do so, could they create a ban that meets the necessary criteria for successful restraint?

623 Work, interview.

Any prohibition would need to clearly distinguish between banned weapons and the many existing weapons that already use autonomy today. It should be possible to clearly differentiate between the kind of defensive human-supervised autonomous weapons that are used today and fully autonomous weapons that would have no human supervision. “Offensive” and “defensive” are distinctions that wouldn’t work, but “fixed” and “mobile” autonomous weapons would. The types of systems in use today are all fixed in place. They are either static (immobile) or affixed to a vehicle occupied by people.

Distinguishing between mobile fully autonomous weapons and advanced missiles would be much harder. The chief difference between the semi-autonomous HARM and the fully autonomous Harpy is the Harpy’s ability to loiter over a wide area and search for targets. Weapons like the Brimstone show how difficult it can be to make this distinction without understanding details about not only the weapon’s functionality, but also its intended use. Distinguishing between recoverable robotic vehicles and non-recoverable munitions would likely be easier.

From the perspective of balancing military necessity against the horribleness of the weapon, these distinctions would be sensible. The most troubling applications of autonomy would be fully autonomous weapons on mobile robotic vehicles. Fixed autonomous weapons would primarily be defensive. By being co-located with human operators, fixed autonomous weapons also would have assured communications so humans could supervise engagements and could physically disable the system if it malfunctioned. Non-recoverable fully autonomous weapons (e.g., loitering munitions) would be permitted, but their risks would be mitigated by the fact that they can’t be sent on patrol. Militaries would want to have some indication that there is an enemy in the vicinity before launching them.

There are other ways nations could draw lines on what is and isn’t allowed, but all these options would all run up against the problem of transparency. How would nations know that others were complying? The United States, Great Britain, France, Russia, China, Israel, and India are already developing experimental stealth drones. Operational versions of these aircraft would be sent into areas where communications might be jammed. Even if nations agreed that these combat drones should not attack targets unless authorized by a human, there would be no way for them to verify others’ compliance. Delegating full autonomy would likely be valuable in some operational
settings. Even if in peacetime nations pledged not to do so, in wartime the temptation might be great enough to overcome any reservations. After all, it’s hard to argue that weapons like the Harpy, TASM, or a radar-hunting combat drone shock the conscience. Using them may entail accepting a different level of risk, but it’s hard to see them as inherently immoral. Further complicating restraint, it might be difficult to even know whether nations were complying with the rules during wartime. If a robot destroyed a target, how would others know whether a human had authorized the target or the robot itself?

All of these factors: clarity, military utility, horribleness of the weapon, and transparency suggest that a ban on fully autonomous weapons is unlikely to succeed. It is almost certain not to pass in the CCW, where consensus is needed, but even if it did, it is hard to see how such rules would remain viable in wartime. Armed robots that had a person in the loop would need only a flip of the switch, or perhaps a software patch, to become fully autonomous. Once a war begins, history suggests that nations will flip the switch, and quickly.

Option 2: Ban anti-personnel autonomous weapons

A ban on autonomous weapons that targeted people may be another matter. The ban is clearer, the horribleness of the weapon greater, and the military utility lower. Transparency is still a problem, but slightly easier in some ways. These factors may make restraint more feasible for anti-personnel autonomous weapons.

It would be easier for states to distinguish between anti-personnel autonomous weapons and existing systems. There are no anti-personnel equivalents of homing missiles or automated defensive systems in use around the world. This could allow states to sidestep the tricky business of carving out exceptions for existing uses.

The balance between military utility and the weapon’s perceived horribleness is also very different for anti-personnel autonomous weapons. Targeting people is much more problematic than targeting objects for a variety of reasons. Complying with IHL is more difficult. Interpreting human behavior and theory-of-mind would likely be required to comply with IHL principles like distinction and hors de combat in many settings. While these challenges might be surmountable someday, current technology does not seem adequate. Anti-personnel autonomous weapons are also significantly

---

625 Precision-guided weapons are evolving down to the level of infantry combat, including some laser-guided munitions such as the DARPA XACTO and Raytheon Spike missile. Because these are laser-guided, they are still remotely controlled by a person.
more hazardous than anti-materiel autonomous weapons. If an anti-personnel weapon malfunctions, humans cannot simply climb out of a tank to escape being targeted. A person can’t stop being human. The potential harm a malfunctioning anti-personnel autonomous weapon could cause would be far greater. Anti-personnel autonomous weapons also pose a greater risk of abuse by those deliberately wanting to attack civilians. They could be used by despots to attack their own population or by countries who don’t respect the law to attack enemy civilians.

Finally, the public may see machines that target and kill people on their own as genuinely horrific. Weapons that autonomously targeted people would tap into an age-old fear of machines rising up against their makers. Like the millennia-old aversion to poison, it may not be logical, but public revulsion could be a decisive factor in achieving political support for a ban.

The military utility of anti-personnel autonomous weapons is also far lower than anti-materiel or anti-vehicle autonomous weapons. The reasons for moving to supervised autonomy (speed) or full autonomy (no communications) don’t generally apply when targeting people. Defensive systems like Aegis need a supervised-autonomous mode to defend against salvos of high-speed missiles, but overwhelming defensive positions through waves of human attackers has not been an effective tactic since the invention of the machine gun. The additional half-second it would take to keep a human in the loop for a weapon like the South Korean sentry gun is marginal. Anti-personnel autonomous weapons in communications-denied environments are also not likely to be of high value for militaries. At the early stages of a war when communications are contested, militaries will be targeting objects such as radars, missile launchers, bases, airplanes, and ships, not people. Militaries would want the ability to use small, discriminating anti-personnel weapons to target specific individuals, such as terrorist leaders, but those would be semi-autonomous weapons; a human would be choosing the target.

Transparency would still be challenging. As is the case for weapons like the South Korean sentry gun, others would have to essentially trust countries when they say they have a human in the loop. Many nations are already fielding armed robotic ground vehicles, and they are likely to become a common feature of future militaries. It would be impossible to verify that these robotic weapons do not have a mode or software patch waiting on the shelf that would enable them to autonomously target people. Given the ubiquity of autonomous technology, it would also be impossible to prevent terrorists from creating homemade autonomous weapons. Large-scale industrial production of anti-personnel weapons would be hard to hide, however. If the military utility on these
weapons were low enough, it isn’t clear that the risk of small-scale uses would compel other nations to violate a prohibition. The combination of low military utility and high potential harm may make restraint possible for anti-personnel autonomous weapons.

Option 3: Establish a “rules of the road” for autonomous weapons

Different problems with autonomous weapons lend themselves to different solutions. A ban on anti-personnel autonomous weapons would reduce the risk of harm to civilians but would not address the problems autonomous weapons pose for crisis stability, escalation control, and war termination. These are very real concerns, and nations will want to cooperate to ensure their robotic systems do not interact in ways that lead to unintended outcomes.

Rather than a treaty, one solution could be to adopt a non-legally binding code of conduct to establish a “rules of the road” for autonomous weapons. The main goal of such a set of rules would be to reduce the potential for unintended interactions between autonomous systems in crises. The best rules would be simple and self-enforcing, like “robotic vehicles should not fire unless fired upon” and “return fire must be limited, discriminating, and proportionate.” The expectation that a robot will not fire unless you shoot first is self-enforcing; it encourages others not to shoot first. Like maritime law, these rules would be intended to govern how autonomous agents interact when they encounter one another in unstructured environments, respecting the right of self-defense but also a desire to avoid unwanted escalation. (Rules governing self-defense against hostile boarding and seizure, like the Chinese seizure of a U.S. underwater drone in the South China Sea in 2016, would be more complicated, but there are any number of rules states could potentially agree on.)

Any rule set could undoubtedly be manipulated by clever adversaries spoiling for a fight, but the main purpose would be to ensure predictable reactions from robotic systems among nations seeking to control escalation. The rules wouldn’t need to be legally binding, since it would be in states’ best interests to cooperate. These rules would likely collapse in war, as submarine rules did, but such rules would at least mitigate the risk of unintended actions during crises.

Option 4: Create a general principle in IHL about the role of human judgment in war

The problem with the above approaches is that technology is always changing. Even the most thoughtful regulations or prohibitions will not be able to foresee all of
the ways that autonomous weapons could evolve over time. An alternative approach would be to focus on the unchanging element in war: the human.626

The laws of war do not specify what role(s) humans should play in lethal force decisions, but perhaps they should. Is there a place for human judgment in war, even if we had all of the technology we could imagine? Should there be limits on what decisions machines make in war, not because they can’t, but because they shouldn’t?

One approach would be to articulate a positive requirement for human involvement in the use of force. In UN discussions, some early consensus has begun to form around the notion that some human involvement is needed in the use of force. This concept has been articulated in different ways, with some NGOs and states calling for “meaningful human control.” The United States, drawing on language in DoD Directive 3000.09, has used the term “appropriate human judgment.” Reflecting these divergent views, the CCW’s final report from its 2016 expert meetings uses the neutral phrase “appropriate human involvement.”627 But no country has suggested that it would be acceptable for there to be no human involvement whatsoever in lethal force decisions. The CCW consensus report from the August 2018 meetings asserted the role of human responsibility explicitly: “Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines.”628 Vague though this statement may be, this common ground is a starting point for cooperation.

Phrases like “meaningful human control,” “appropriate human judgment,” “effective human control,” and “appropriate human involvement” all seem to get at this

---


While these terms are not yet defined, they suggest broad agreement that there is some irreducible role for humans in lethal force decisions on the battlefield. Setting aside for the moment the specific label, what would be the underlying idea behind a principle of “______ human ______”? IHL may help give us some purchase on the problem, if one adopts the viewpoint that the laws of war apply to people, not machines. This is the view captured in the U.S. DoD Law of War Manual:

The law of war rules on conducting attacks (such as the rules relating to discrimination and proportionality) impose obligations on persons. These rules do not impose obligations on the weapons themselves; … Rather, it is persons who must comply with the law of war.

If it is true that humans are obligated under IHL to make a determination about the lawfulness of an attack and they cannot to delegate this obligation to a machine, then certain things flow from that position.

In order for a human to make a determination about whether an attack will comply with the principles of distinction, proportionality, and precautions in attack, the human would need to have some minimum amount of information about the specific attack. The human must have sufficient information about the target(s), the weapon, the environment, and the context for attack in order to determine whether that particular

---


attack is lawful. It also means that the attack must be bounded in time, space, targets, and means of attack in order for the determination about the lawfulness of that attack to be meaningful. There would presumably be some conditions (time elapsed, geographic boundaries crossed, circumstances changed) under which the human’s determination about the lawfulness of the attack might no longer be valid.

How much information the person needs and what those bounds are on autonomy is open for debate. This perspective would seem to suggest, though, that IHL requires some minimum necessary level of human involvement in lethal force decisions: (1) human judgment about the lawfulness of an attack; (2) some specific information about the target(s), weapon, environment, and context for attack in order to make a determination about lawfulness of that particular attack; and (3) that the weapon’s autonomy be bounded in space, time, possible targets, and means of attack.

There may be other ways of phrasing this principle and reasonable people might disagree with this interpretation, but there could be merit in countries reaching agreement on a common standard for necessary human involvement in lethal force. While an overarching principle along these lines would not directly tell states which weapons are permitted and which are not, it could be a common starting point for evaluating technology as it evolves. Many principles in IHL are open to interpretation: unnecessary suffering, proportionality, and precautions in attack, for example. These terms do not tell states which weapons cause unnecessary suffering or how much collateral damage is proportionate, but they still have value. Similarly, a broad principle outlining the role of human judgment in war could be a valuable benchmark against which to evaluate future weapons.

Reframing the debate

Humanity is at the threshold of a new technology that could fundamentally change our relationship with war. It is an encouraging sign that nations have come together through the United Nations CCW to discuss autonomous weapons, but to-date strategic considerations have gotten short shrift. A few experts have presented on offense-defense balance and arms races, but there has been virtually no discussion of how autonomous weapons might complicate crisis stability, escalation control, and war termination. John Borrie from the UN Institute for Disarmament Research is concerned

---

about the risk of “unintended lethal effects” from autonomous weapons, but he acknowledged, “it's not really a significant feature of the policy debate in the CCW."

This is unfortunate, because autonomous weapons raise important issues for stability. Supervised autonomous weapons have been used relatively safely for decades, but fully autonomous weapons would be a novel development and a fundamental paradigm shift in warfare. In deploying fully autonomous weapons, militaries would be introducing onto the battlefield a weapon designed to engage targets that have not been authorized by a human and which they cannot control or recall once launched. They would be sending this weapon into an environment that they do not control where it is subject to enemy hacking and manipulation. Nothing in the laws of war suggests that this is inherently illegal. It undoubtedly introduces a new element of hazard in war, however. It would be facile and wrong to suggest that fully autonomous weapons would, overall, be safer and more humane than semi-autonomous weapons that retain a human in the loop. This argument conflates the benefits of adding autonomy, which are significant, with completely removing the human from the loop. There may be military operational benefits to fully autonomous weapons. They would allow deep strikes inside contested areas against mobile and relocatable targets. These are vitally important for militaries, and the ability to conduct these strikes with robotic systems without communications links would unquestionably be valuable. One can even imagine cases where their use would result in more humane outcomes, provided they functioned properly, such as hostage rescue in communications-denied environments or destroying mobile missiles launchers armed with WMD. On the whole, though, the net effect of introducing fully autonomous weapons on the battlefield is likely to be increased speed, greater hazard if accidents occur, and reduced human control. States have every incentive to cooperate to avoid a world where they have less control over the use of force. Mutual restraint is definitely in states’ interests. This is especially true for great powers, given the destruction that war among them would bring.

Diplomats and members of civil society who are involved in discussions on autonomous weapons would maximize their chances of achieving a constructive solution by reframing the debate in a few significant ways. First, stability should be a bigger focus of discussions. Autonomous weapons pose serious risks to stability. These risks have not been a significant focus of international discussions on autonomous weapons to-date, yet they are an important factor that states should consider when

---

632 Borrie, interview.
evaluating these weapons. Second, any discussions surrounding regulations must seriously grapple with the challenge of transparency. States will be reluctant to give up weapons if they cannot verify through some means that others are doing so as well. Trust among nations is a major obstacle to a successful agreement. Fears by states that others might cheat and gain an advantage are real and valid, and these concerns must be addressed for any meaningful agreement to work. Third, discussions need to be reframed to make an argument that would be in powerful states’ best interests to support a solution. Major military powers will need to be a party to any regulation in order for it to be successful, and they need to see why it is in their self-interest to do so. Fourth, diplomats will need to make progress on working definitions, if nothing else to ensure that discussants are talking about the same concepts. And finally, discussions should build on areas of agreement. States have begun to make progress on areas like human responsibility. This should be further explored.

States have a window of opportunity to make progress on mitigating some of the risks surrounding autonomous weapons. While there are challenges in preemptively regulating weapons because the shape of the technology is not yet fully known, there are also problems in attempting to limit technologies once they are widespread. States have a window of opportunity in the next few years to put in place commonsense rules surrounding autonomous weapons that can help mitigate some of their dangers.
CONCLUSION

The military robotics revolution is well underway. Militaries around the globe are building ever more sophisticated robotic vehicles, munitions, and cyber weapons and across all domains – land, air, sea, and cyber. With each generation these systems incorporate more autonomy. Where this trajectory is headed remains an open question. No nation has explicitly stated they intend to build fully autonomous weapons that could search for, select, and engage targets on their own. However, the course of technological development will make such weapons possible, including by non-state actors. Simple autonomous weapons are already possible with existing technology. A few weapons, like the Israeli Harpy, cross the line to fully autonomous weapons but they are not in widespread use today. Supervised autonomous weapons, which have a human “on the loop,” are widely used but only for limited defensive purposes. A number of next-generations weapons, such as the UK Taranis or DARPA CODE program, come right up to the line of full autonomy.

Many countries have been ambiguous about whether they intend to cross the line to fully autonomous weapons. Even strong statements about keeping humans “in the loop,” such as those from the United States and United Kingdom, appear more nuanced at close inspection. Other major military powers, such as Russia and China, have been far less clear with their intentions. And while twenty-seven states have said they support a ban on autonomous weapons, none are major military powers or robotics developers. The capability to build simple fully autonomous weapons is widespread today and over time will only become more advanced. Widespread deployment is a possibility. At present, despite significant international debate and five years of discussions at the United Nations, there are no meaningful barriers to developing and deploying fully autonomous weapons, and no major military power or robotics developer has definitively renounced their use.

633 Campaign to Stop Killer Robots, “Country Views on Killer Robots.”

The introduction of fully autonomous weapons that could search for, select, and engage targets on their own would be a novel development in warfare. While this would be a logical next step in a long trajectory of increasing autonomy and automation in weapons, dating back to the Gatling gun, it would introduce a qualitatively different paradigm for humans’ relationship with violence. While autonomous weapons would be built, programmed, and set into operation by humans, humans would no longer be in control of the specific target being engaged. The possible consequences of such a development have been extensively explored from a number of dimensions – legal, ethical, and moral – but there has been little analysis to-date of their implications on stability. A number of scholars – Michael Carl Haas, Alex Velez-Green, Jean-Marc Rickli, Frank Sauer, Jürgen Altmann, Kenneth Payne, John Borrie, and Michael Horowitz – have raised concerns about the implications of autonomous weapons on stability. The issue has been most thoroughly examined by Frank Sauer and Jürgen Altmann. Yet key questions remain, making this an underexplored area worthy of further analysis.

In Part III, Stability and Autonomous Weapons, this paper explored a number of ways that autonomous weapons could be potentially stabilizing or destabilizing. A major factor across many of these hypothesized effects on stability is the extent to which autonomous weapons increase or decrease human control over lethal actions on the battlefield. The nature of autonomy is such that it entails the delegation of a task to a machine. When the machine performs that task reliably, consistent with human intentions, autonomy can often increase human control over a task. A thermostat, for example, enables a more consistent room temperature than if a human had to manually turn on and off heat or air conditioning to achieve the desired temperature. However, if an autonomous system begins performing its task in an erroneous manner, for any number of reasons, then human control is reduced over the task. For autonomous weapons, the fear is that erroneous action could lead to unintended lethal engagements that escalate a crisis or conflict or make it more difficult to terminate conflict. On the other hand, the potential benefit of autonomous weapons from the standpoint of stability is that, by replacing human decision-making with automation, autonomous weapons would more directly link political leaders to tactical actions. In principle, this could improve stability in a crisis by reducing the potential for unauthorized actions for subordinates that could be escalatory or could lead adversaries to misperceive a state’s intentions. This led to two conflicting hypotheses about how autonomous weapons could affect stability:
H1: Widespread autonomous weapons would undermine crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could initiate, escalate, and/or make it more difficult to terminate conflicts.

The alternative hypothesis was:

H2: Widespread autonomous weapons would improve crisis stability, escalation control, and war termination by increasing political leaders’ control over their military forces, reducing the risk of accidents or unauthorized escalatory actions.

These hypotheses were tested in Part IV, Controlling Autonomous Weapons: Risk, Predictability, and Hazard, by examining over 25 case studies in military and non-military autonomous systems. These case studies were intended to better understand whether fully autonomous weapons would be likely to function as intended in conflict or crisis settings.

To better understand their likely reliability in real-world conditions, Part IV brought to bear existing theories of accidents and reliability in complex systems, the competing theories Normal Accident Theory and High-Reliability Organizations. Part IV concluded that complex systems are inherently susceptible to failure because of the inability of human designers and testers to fully anticipate all of the possible interactions between the system and its environment, its human operators, and even within elements of the system itself. In loosely coupled systems, in which one action does not immediately and necessarily lead to another action, such as in human bureaucracies, there is often an opportunity for the system to adjust to failures by adapting processes and procedures to accomplish a task. In tightly coupled systems, where one action immediately leads to another action with little flexibility for deviation of procedures, failures can cascade into accidents in which the system is unable to adapt. Normal accident theory predicts that accidents are “normal” in complex, tightly-coupled systems – that is, they are expected features of the system’s behavior. Normal accident theory predicts that increased testing can reduce the probability of accidents, but that they can never be eliminated. Conversely, the theory of high-reliability organizations predicts that some organizations that have a commitment to safety can, in fact, achieve exceptionally low accident rates even when operating complex, hazardous systems. The U.S. FAA air traffic control system – or more broadly U.S. airline safety –
is one example of high-reliability in action. Despite carrying billions of passengers across 100 million flights per year, U.S. air carriers have suffered only a single passenger fatality in nine years, an astonishing safety record.\textsuperscript{635} While militaries as a whole would not be considered high-reliability organizations, there some examples of specific military communities that have achieved high reliability. The U.S. Navy’s submarine community, for example, has not lost a SUBSAFE-certified submarine since instituting the program in 1963.

Automation plays a mixed role in accidents. Automation can be extremely valuable in improving safety and reliability for routine operations and has played a major factor in increasing airline safety for example. On the other hand, in novel situations automation often is “brittle” and lacks the flexibility to adapt, which can lead to accidents. A number of case studies were examined of autonomous systems in both military and non-military settings, from stock trading to missile defense. The conclusion from these case studies is that high-reliability operation of complex, autonomous systems is possible when a few conditions are met. These include routine, day-to-day operation of the system in its normal operating environment and the presence of highly skilled and experienced human operators, who are tightly integrated with the system. The peacetime performance of the U.S. Navy’s Aegis weapon system is an example of such a system. When these conditions are not met, normal accident theory more appropriately explains the likely accident rates of such systems. Examples include the 2003 Patriot fratricides and the Aegis’ performance in the 1988 Persian Gulf crisis with Iran, in which the USS \textit{Vincennes} shot down an Iranian airliner. More automation in some settings may help improve safety, and indeed better automated safeties might have prevented the \textit{Vincennes} incident. But in other settings, greater automation can introduce complexity that may lead to unexpected interactions between the system and its environment, within the system itself, or confusion by human operators trying to understand the systems’ behavior. Numerous automation-related stock trading accidents, including “flash crashes,” are examples of situations where automation has contributed to causing or exacerbating accidents.

For fully autonomous weapons, several conditions exist that are likely to undermine reliability and make high-reliability unlikely. These include complexity, tight-coupling, the presence of adversarial actors trying to undermine the system and

\textsuperscript{635} Bloomberg, “‘We Shouldn’t Stop Flying.’ What to Make of the Southwest Plane Accident That Shattered a Remarkable U.S. Safety Record,” April 18, 2018, \url{http://fortune.com/2018/04/18/southwest-plane-accident-safety-record/}. 
induce failures, novel environmental conditions in wartime or crisis settings, and the lack of human operators who can adjust to new situations or adversarial adaptation. This suggests that accident rates for fully autonomous weapons are most likely to align with the predictions of normal accident theory. Better testing can improve reliability to a degree, but high-reliability is unlikely. If fully autonomous weapons were deployed, accidents would be expected, if rare. From the standpoint of stability, this would lend support to hypothesis 1: “Widespread autonomous weapons would undermine crisis stability, escalation control, and war termination because of the potential for unintended lethal engagements that could initiate, escalate, and/or make it more difficult to terminate conflicts.”

Advances in artificial intelligence are not likely to fundamentally change this dynamic. Machine learning systems that learn their behavior, rather than follow a strict set of if-then rules, are more capable at many tasks than rule-based systems but are even more challenging from a reliability standpoint. More advanced forms of artificial intelligence, such as hypothesized future human-like artificial general intelligence, would be able to adapt to novel situations and understand context, but would introduce other, potentially even more serious, control problems. While AI researchers are working on improving AI safety, no breakthrough appears on the horizon that would fundamentally change the risk of autonomous systems performing erroneously in novel, adversarial situations without human involvement.

An examination of the military operational conditions under which autonomous weapons might be used suggests that, while some of the military advantages of fully autonomous weapons might be overstated, they nevertheless are likely to be useful in some discrete settings. Should accidents occur, fully autonomous weapons could cause unintended lethal engagements that result in fratricide, civilian casualties, or unauthorized against adversary forces. In some settings, these could complicate crises or escalate tensions, suggesting that concerns about autonomous weapons undermining stability are valid. In physical space, escalation to a “flash war” that unfolds rapidly before humans can effectively intervene is unlikely simply due to the physical constraints of time and space on transiting missiles, aircraft, ships, or other platforms. Escalation could occur much faster in cyberspace. Even in physical space, unintended lethal engagements from autonomous weapons could escalate tensions to the point where it becomes difficult from policymakers to de-escalate a conflict, or an incident could be used as a casus belli by leaders who desire war.

A number of policy options for addressing autonomous weapons have been proposed, ranging from a preemptive legally-binding treating banning such weapons to
simply relying on existing international law to guide their development. Part VII considered four policy options: (1) a comprehensive ban on autonomous weapons; (2) a more limited ban on anti-personnel autonomous weapons; (3) a “rules of the road” for how autonomous weapons interact in peacetime or crisis settings; and (4) establishing a general principle in the laws of war about the role of human judgment. Based on historical experience with weapons bans throughout history, Part VII suggests that a comprehensive ban is unlikely to succeed for a number of factors and that the other policy options have a greater likelihood of success, although success is far from assured for any of them and each of them address more narrow concerns. For example, a ban on anti-personnel autonomous weapons most directly addresses concerns about civilian harm. A “rules of the road” for autonomous weapons most directly addresses concerns about crisis stability and unintended escalation. These predictions have value from an academic standpoint; they can be tested as development of autonomous weapons or other weapons technologies proceed. They are also of value to policymakers seeking to adapt to the reality of the technological feasibility of autonomous weapons and find policy solutions that mitigate the worst outcomes.

An analysis of the possible effects of autonomous weapons on stability finds that widespread deployment of fully autonomous weapons would likely undermine crisis stability, escalation control, and war termination because of the risk of unintended lethal engagements. These issues are worthy of further academic study, and policymakers should ensure that stability is considered in ongoing international discussions on lethal autonomous weapons.
# APPENDICES

## A. Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>anti-aircraft artillery</td>
</tr>
<tr>
<td>ABM</td>
<td>Anti-Ballistic Missile</td>
</tr>
<tr>
<td>ACTUV</td>
<td>Anti-submarine warfare Continuous Trail Unmanned Vessel</td>
</tr>
<tr>
<td>AGI</td>
<td>artificial general intelligence</td>
</tr>
<tr>
<td>AGM</td>
<td>air-to-ground missile</td>
</tr>
<tr>
<td>AI</td>
<td>artificial intelligence</td>
</tr>
<tr>
<td>AMRAAM</td>
<td>Advanced Medium-Range Air-to-Air Missile</td>
</tr>
<tr>
<td>ARPA</td>
<td>Advanced Research Projects Agency</td>
</tr>
<tr>
<td>ASI</td>
<td>artificial superintelligence</td>
</tr>
<tr>
<td>ASW</td>
<td>anti-submarine warfare</td>
</tr>
<tr>
<td>ATR</td>
<td>automatic target recognition</td>
</tr>
<tr>
<td>BDA</td>
<td>battle damage assessment</td>
</tr>
<tr>
<td>BWC</td>
<td>Biological Weapons Convention</td>
</tr>
<tr>
<td>CCW</td>
<td>Convention on Certain Conventional Weapons</td>
</tr>
<tr>
<td>C&amp;D</td>
<td>Command and Decision</td>
</tr>
<tr>
<td>CIC</td>
<td>combat information center</td>
</tr>
<tr>
<td>CIWS</td>
<td>Close-In Weapon System</td>
</tr>
<tr>
<td>CODE</td>
<td>Collaborative Operations in Denied Environments</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>DDoS</td>
<td>distributed denial of service</td>
</tr>
<tr>
<td>DIY</td>
<td>do-it-yourself</td>
</tr>
<tr>
<td>DMZ</td>
<td>demilitarized zone</td>
</tr>
<tr>
<td>DoD</td>
<td>Department of Defense</td>
</tr>
<tr>
<td>FAA</td>
<td>Federal Aviation Administration</td>
</tr>
<tr>
<td>FIAC</td>
<td>fast inshore attack craft</td>
</tr>
<tr>
<td>FIS</td>
<td>Fire Inhibit Switch</td>
</tr>
<tr>
<td>FLA</td>
<td>Fast Lightweight Autonomy</td>
</tr>
<tr>
<td>GGE</td>
<td>Group of Governmental Experts</td>
</tr>
<tr>
<td>GPS</td>
<td>global positioning system</td>
</tr>
<tr>
<td>ICRAC</td>
<td>International Committee for Robot Arms Control</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>ICRC</td>
<td>International Committee of the Red Cross</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>IFF</td>
<td>identification friend or foe</td>
</tr>
<tr>
<td>IHL</td>
<td>international humanitarian law</td>
</tr>
<tr>
<td>IMU</td>
<td>inertial measurement unit</td>
</tr>
<tr>
<td>INF</td>
<td>Intermediate-Range Nuclear Forces</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>J-UCAS</td>
<td>Joint Unmanned Combat Air Systems</td>
</tr>
<tr>
<td>LIDAR</td>
<td>light detection and ranging</td>
</tr>
<tr>
<td>LOCAAS</td>
<td>Low Cost Autonomous Attack System</td>
</tr>
<tr>
<td>LRASM</td>
<td>Long-Range Anti-Ship Missile</td>
</tr>
<tr>
<td>MAD</td>
<td>mutual assured destruction</td>
</tr>
<tr>
<td>MARS</td>
<td>Mobile Autonomous Robotic System</td>
</tr>
<tr>
<td>MMW</td>
<td>millimeter-wave</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
</tr>
<tr>
<td>NGO</td>
<td>nongovernmental organization</td>
</tr>
<tr>
<td>NORAD</td>
<td>North American Aerospace Defense Command</td>
</tr>
<tr>
<td>ONR</td>
<td>Office of Naval Research</td>
</tr>
<tr>
<td>OODA</td>
<td>observe, orient, decide, act</td>
</tr>
<tr>
<td>OPM</td>
<td>Office of Personnel Management</td>
</tr>
<tr>
<td>PGM</td>
<td>precision-guided munition</td>
</tr>
<tr>
<td>PLC</td>
<td>programmable logic controllers</td>
</tr>
<tr>
<td>RAS</td>
<td>IEEE Robotics and Automation Society</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>research and development</td>
</tr>
<tr>
<td>ROE</td>
<td>rules of engagement</td>
</tr>
<tr>
<td>SAG</td>
<td>surface action group</td>
</tr>
<tr>
<td>SAR</td>
<td>synthetic aperture radar</td>
</tr>
<tr>
<td>SAW</td>
<td>Squad Automatic Weapon</td>
</tr>
<tr>
<td>SEC</td>
<td>Securities and Exchange Commission</td>
</tr>
<tr>
<td>SFW</td>
<td>Sensor Fuzed Weapon</td>
</tr>
<tr>
<td>SORT</td>
<td>Strategic Offensive Reductions Treaty</td>
</tr>
<tr>
<td>START</td>
<td>Strategic Arms Reduction Treaty</td>
</tr>
<tr>
<td>SUBSAFE</td>
<td>Submarine Safety</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>TASM</td>
<td>Tomahawk Anti-Ship Missile</td>
</tr>
<tr>
<td>TBM</td>
<td>tactical ballistic missile</td>
</tr>
<tr>
<td>TJ</td>
<td>Thomas Jefferson High School</td>
</tr>
<tr>
<td>TLAM</td>
<td>Tomahawk Land Attack Missile</td>
</tr>
<tr>
<td>TRACE</td>
<td>Target Recognition and Adaption in Contested Environments</td>
</tr>
<tr>
<td>TTO</td>
<td>Tactical Technology Office</td>
</tr>
<tr>
<td>TTP</td>
<td>tactics, techniques, and procedures</td>
</tr>
<tr>
<td>UAV</td>
<td>uninhabited aerial vehicle</td>
</tr>
<tr>
<td>UCAS</td>
<td>unmanned combat air system</td>
</tr>
<tr>
<td>UCAV</td>
<td>uninhabited combat aerial vehicle</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>UNIDIR</td>
<td>UN Institute for Disarmament Research</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>WMD</td>
<td>weapons of mass destruction</td>
</tr>
</tbody>
</table>
B. Definitions of Autonomous Weapons

The U.S. Department of Defense (DoD) official policy on autonomy in weapon systems, DoD Directive 3000.09, Autonomy in Weapon Systems, uses the following definitions.636

*autonomous weapon system:* A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.

*human-supervised autonomous weapon system:* An autonomous weapon system that is designed to provide human operators with the ability to intervene and terminate engagements, including in the event of a weapon system failure, before unacceptable levels of damage occur.

*semi-autonomous weapon system:* A weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator. This includes:

Semi-autonomous weapon systems that employ autonomy for engagement-related functions including, but not limited to, acquiring, tracking, and identifying potential targets; cueing potential targets to human operators; prioritizing selected targets; timing of when to fire; or providing terminal guidance to home in on selected targets, provided that human control is retained over the decision to select individual targets and specific target groups for engagement.

“Fire and forget” or lock-on-after-launch homing munitions that rely on tactics, techniques, and procedures to maximize the probability that the only targets within the seeker’s acquisition basket when the seeker activates are those individual targets or specific target groups that have been selected by a human operator.

---

**target selection:** The determination that an individual target or a specific group of targets is to be engaged.

Christof Heyns, UN Special Rapporteur on extrajudicial, summary or arbitrary executions, wrote in his April 2013 report:  

The measure of autonomy that processors give to robots should be seen as a continuum with significant human involvement on one side, as with UCAVs where there is “a human in the loop”, and full autonomy on the other, as with [lethal autonomous robotics (LARs)] where human beings are “out of the loop.

Under the currently envisaged scenario, humans will at least remain part of what may be called the “wider loop”: they will programme the ultimate goals into the robotic systems and decide to activate and, if necessary, deactivate them, while autonomous weapons will translate those goals into tasks and execute them without requiring further human intervention.

Supervised autonomy means that there is a “human on the loop” (as opposed to “in” or “out”), who monitors and can override the robot’s decisions. However, the power to override may in reality be limited because the decision-making processes of robots are often measured in nanoseconds and the informational basis of those decisions may not be practically accessible to the supervisor. In such circumstances humans are de facto out of the loop and the machines thus effectively constitute LARs.

“Autonomous” needs to be distinguished from “automatic” or “automated.” Automatic systems, such as household appliances, operate within a structured and predictable environment. Autonomous systems can function in an open environment, under unstructured and dynamic circumstances. As such their actions (like those of humans) may ultimately be unpredictable, especially in

---

situations as chaotic as armed conflict, and even more so when they interact with other autonomous systems.

The terms “autonomy” or “autonomous”, as used in the context of robots, can be misleading. They do not mean anything akin to “free will” or “moral agency” as used to describe human decision-making. Moreover, while the relevant technology is developing at an exponential rate, and full autonomy is bound to mean less human involvement in 10 years’ time compared to today, sentient robots or strong artificial intelligence are not currently in the picture.

Human Rights Watch, in their 2012 report “Losing Humanity,” used the following terminology: 638

Although experts debate the precise definition, robots are essentially machines that have the power to sense and act based on how they are programmed. They all possess some degree of autonomy, which means the ability of a machine to operate without human supervision. The exact level of autonomy can vary greatly. Robotic weapons, which are unmanned, are often divided into three categories based on the amount of human involvement in their actions:

• **Human-in-the-Loop Weapons**: Robots that can select targets and deliver force only with a human command;

• **Human-on-the-Loop Weapons**: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots’ actions; and

• **Human-out-of-the-Loop Weapons**: Robots that are capable of selecting targets and delivering force without any human input or interaction.

---

The International Committee of the Red Cross (ICRC), in a working paper submitted to the CCW in 2016, stated:639

For the purpose of better understanding the key issues, the ICRC understands autonomous weapon systems to be: "Any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention."

---

C. Autonomous Weapons in Use Today: Selected Examples

*Examples of Fully Autonomous Weapons: Loitering Munitions*

**HARPY**

<table>
<thead>
<tr>
<th><strong>Other Name</strong></th>
<th>Harpy NG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date of Introduction</strong></td>
<td>1990-1991</td>
</tr>
<tr>
<td><strong>Manufacturer</strong></td>
<td>Israel Aerospace Industries</td>
</tr>
<tr>
<td><strong>Manufacturing Country</strong></td>
<td>Israel</td>
</tr>
<tr>
<td><strong>National Operators</strong></td>
<td>Chile, China, India, Israel, South Korea, Turkey</td>
</tr>
</tbody>
</table>

**References**


**HAROP**

<table>
<thead>
<tr>
<th><strong>Other Name</strong></th>
<th>Harpy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date of Introduction</strong></td>
<td>2005</td>
</tr>
<tr>
<td>Manufacturer</td>
<td>Israel Aerospace Industries</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Manufacturing Country</td>
<td>Israel</td>
</tr>
<tr>
<td>National Operators</td>
<td>Azerbaijan, India, Israel</td>
</tr>
</tbody>
</table>

**TOMAHAWK ANTI-SHIP MISSILE (TASM)**

<table>
<thead>
<tr>
<th>Other Name</th>
<th>BGM/RGM-109B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of Introduction</td>
<td>1982</td>
</tr>
<tr>
<td>Retired from Service</td>
<td>1994</td>
</tr>
<tr>
<td>Manufacturer</td>
<td>General Dynamics</td>
</tr>
</tbody>
</table>
Examples of Fully Autonomous Weapons:

Encapsulated Torpedo Mines

CAPTOR

Other Name  Mk 60 enCAPsulated TORpedo mine
Date of Introduction  1979 (production stopped in 1986)
Manufacturer  Goodyear Aerospace
Manufacturing Country  United States
National Operators  United States
References

PMK-2 MINE

Date of Introduction  Unknown
<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing Country</td>
<td>Russia</td>
</tr>
<tr>
<td>National Operators</td>
<td>Russia, China</td>
</tr>
</tbody>
</table>

**References**


Selected Examples of Human-Supervised Autonomous Weapons:

*Air and Missile Defense Systems*

**AEGIS COMBAT SYSTEM**

<table>
<thead>
<tr>
<th>Date of Introduction</th>
<th>1983</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer</td>
<td>Lockheed Martin</td>
</tr>
<tr>
<td>Manufacturing Country</td>
<td>United States</td>
</tr>
<tr>
<td>National Operators</td>
<td>Australia, Japan, Norway, South Korea, Spain, and the United States</td>
</tr>
</tbody>
</table>

**References**


Lockheed Martin, “Aegis Combat System.”


### GOALKEEPER

<table>
<thead>
<tr>
<th>Date of Introduction</th>
<th>Early 1990s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer</td>
<td>Thales Netherlands</td>
</tr>
<tr>
<td>Manufacturing Country</td>
<td>Netherlands</td>
</tr>
<tr>
<td>National Operators</td>
<td>Belgium, Chile, the Netherlands, Portugal, Qatar, South Korea, the United Arab Emirates, and the United Kingdom</td>
</tr>
</tbody>
</table>

**References**


### KASHTAN CLOSE-IN WEAPON SYSTEM

<table>
<thead>
<tr>
<th>Date of Introduction</th>
<th>1988</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer</td>
<td>KBP Instrument Design Bureau</td>
</tr>
<tr>
<td>Manufacturing Country</td>
<td>Russia</td>
</tr>
<tr>
<td>National Operators</td>
<td>China, India, Russia</td>
</tr>
</tbody>
</table>

**References**


MK 15 PHALANX CLOSE-IN WEAPON SYSTEM

Date of Introduction 1980

Manufacturer The Raytheon Company

Manufacturing Country United States

National Operators 25 navies including Australia, Bahrain, Canada, Egypt, India, Israel, Japan, New Zealand, Pakistan, Poland, Portugal, Saudi Arabia, South Korea, Taiwan, the United Kingdom, and the United States

References

The Raytheon Company, “Phalanx Close-In Weapon System (CIWS).”
“Raytheon services India's Phalanx,” UPI, July 22, 2010.
“Raytheon, South Korea in Phalanx deal,” UPI, February 25, 2014.

PATRIOT

Date of Introduction 1984

Manufacturer The Raytheon Company

Manufacturing Country United States
National Operators

Egypt, Germany, Greece, Israel, Japan, Kuwait, the Netherlands, Saudi Arabia, South Korea, Spain, Taiwan, the United Arab Emirates, and the United States

References

The Raytheon Company, “Patriot.”

SeaRAM ANTI-SHIP MISSILE DEFENSE SYSTEM

Date of Introduction

2008

Manufacturer

The Raytheon Company

Manufacturing Country

United States

National Operators

Japan and the United States

References

Selected Examples of Human-Supervised Autonomous Weapons:  
Ground Vehicle Active Protection Systems

AMAP-ADS

Date of Introduction 2011

Manufacturer IBD Deisenroth Engineering

Manufacturing Country Germany

National Operators “An Asian nation” (rumored to be Singapore)

References
Rheinmetall Defence, “Rheinmetall takes up a majority share in ADS GmbH,” January 2, 2011.

IBD Deisenroth Engineering, “AMAP™-ADS: The Active Protection System.”

Rheinmetall Defence, “Rheinmetall takes up a majority share in ADS GmbH.”


ARENA

Date of Introduction 1993

Manufacturer KBP Instrument Design Bureau

Manufacturing Country Russia

National Operators Russia

References

“ARENA Active Protection System.”

Meyer, “Active Protection Systems: Impregnable Armor or Simply Enhanced Survivability?”
DROZD

Date of Introduction 1983

Manufacturer KBP Instrument Design Bureau

Manufacturing Country Russia

National Operators Russia

References


DROZD-2

Date of Introduction 2005

Manufacturer KBP Instrument Design Bureau

Manufacturing Country Russia

National Operators Russia

References


IRON CURTAIN

Date of Introduction In development

Manufacturer Artis, LLC

Manufacturing Country United States

National Operators United States
References
David Hambling, “New Bomb-Resistant Trucks Will Blast RPGs Before They Hit,”
WIRED, November 27, 2009.
Artis, LLC, “Iron Curtain.”
Hambling, “New Bomb-Resistant Trucks Will Blast RPGs Before They Hit.”

LEDS-150

Date of Introduction 2015 (anticipated)
Manufacturer Saab
Manufacturing Country South Africa
National Operators Pakistan and India are reportedly interested in purchasing the LEDS-150

References
Saab, “LEDS full spectrum active protection for land vehicles.”
Martin, “Saab seeking LEDS customer.”

QUICK KILL

Date of Introduction In development
Manufacturer The Raytheon Company
Manufacturing Country United States
National Operators United States

References
The Raytheon Company, “Raytheon's Quick Kill Active Protection System defeats one of the most lethal armor-piercing Rocket Propelled Grenades,” January 9, 2013.
The Raytheon Company, “Quick Kill™ Active Protection System (APS).”

SHARK

Date of Introduction 2006 (est.)

Manufacturer Thales Group and IBD Deisenroth Engineering

Manufacturing Country France

National Operators France

References

TRENCH COAT

Date of Introduction 2012 (est.)

Manufacturer Rafael Advanced Defensive Systems

Manufacturing Country Israel

National Operators Israel

References
Amos Harel, “Precise missiles, cyber-attacks, mass border invasions: Israel's next war,” Haaretz, October 12, 2013.
TROPHY

Other Name: ASPRO-A

Date of Introduction: 2007

Manufacturer: Rafael Advanced Defensive Systems

Manufacturing Country: Israel

National Operators: Israel

References

“Trophy Active Protection System,” Defense Update.

Rafael Advanced Defensive Systems, “Trophy.”
### D. Historical Weapons Bans

<table>
<thead>
<tr>
<th>Era</th>
<th>Weapon</th>
<th>Year</th>
<th>Regulation or Treaty</th>
<th>Legally binding?</th>
<th>Type of Regulation</th>
<th>Successful?</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Modern Era</td>
<td>poisoned or barbed arrows</td>
<td>Dates vary - 1500 to 200 BC</td>
<td>Laws of Manu; Dharmaśāstras; Mahābhārata</td>
<td>legally binding</td>
<td>banned use</td>
<td>success unknown</td>
<td>unnecessary suffering</td>
</tr>
<tr>
<td></td>
<td>concealed weapons</td>
<td>Dates vary - 1500 to 200 BC</td>
<td>Laws of Manu</td>
<td>legally binding</td>
<td>banned use</td>
<td>success unknown</td>
<td>perfidy</td>
</tr>
<tr>
<td></td>
<td>fire-tipped weapons</td>
<td>Dates vary - 1500 to 200 BC</td>
<td>Laws of Manu</td>
<td>legally binding</td>
<td>banned use</td>
<td>success unknown</td>
<td>unnecessary suffering</td>
</tr>
<tr>
<td></td>
<td>crossbow</td>
<td>1097; 1139</td>
<td>1097 Lateran Synod; 1139 Second Lateran Council</td>
<td>legally binding</td>
<td>banned use</td>
<td>failed</td>
<td>political control</td>
</tr>
<tr>
<td></td>
<td>firearms</td>
<td>1607-1867</td>
<td>Tokugawa Shogunate Japan</td>
<td>legally binding</td>
<td>effectively prohibited production</td>
<td>successful (lasted ~250 years)</td>
<td>political control</td>
</tr>
<tr>
<td></td>
<td>firearms</td>
<td>1523-1543</td>
<td>King Henry VIII</td>
<td>legally binding</td>
<td>limited ownership among civilian population</td>
<td>short-lived</td>
<td>political control</td>
</tr>
<tr>
<td></td>
<td>poisoned bullets</td>
<td>1675</td>
<td>Strasbourg Agreement</td>
<td>legally binding</td>
<td>banned use</td>
<td>successful</td>
<td>unnecessary suffering</td>
</tr>
<tr>
<td>Turn of the Century</td>
<td>1868</td>
<td>1868 St. Petersburg Declaration</td>
<td>legally binding</td>
<td>banned use</td>
<td>superceded by technology, but adhered to in spirit</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-------</td>
<td>---------------------------------</td>
<td>----------------</td>
<td>------------</td>
<td>-----------------------------------------------------</td>
<td>----------------------</td>
<td></td>
</tr>
<tr>
<td>explosive or inflammable projectiles below 400 grams</td>
<td>1868</td>
<td>1868 St. Petersburg Declaration</td>
<td>legally binding</td>
<td>banned use</td>
<td>superceded by technology, but adhered to in spirit</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>expanding bullets</td>
<td>1899</td>
<td>1899 Hague Declaration</td>
<td>legally binding</td>
<td>banned use</td>
<td>successful in limiting battlefield use, although lawful in civilian applications</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>asphyxiating gases (from projectiles)</td>
<td>1899</td>
<td>1899 Hague Declaration</td>
<td>legally binding</td>
<td>banned use</td>
<td>failed - used in WW I</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>poison</td>
<td>1899; 1907</td>
<td>1899 and 1907 Hague Declarations</td>
<td>legally binding</td>
<td>banned use</td>
<td>successful</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>weapons that cause superfluous injury</td>
<td>1899; 1907</td>
<td>1899 and 1907 Hague Declarations</td>
<td>legally binding</td>
<td>banned use</td>
<td>mixed, but generally successful</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>balloon-delivered projectiles or explosives</td>
<td>1899; 1907</td>
<td>1899 and 1907 Hague Declarations</td>
<td>legally binding</td>
<td>banned use</td>
<td>short-lived</td>
<td>civilian casualties</td>
<td></td>
</tr>
<tr>
<td>aerial bombardment against undefended cities</td>
<td>1907</td>
<td>1907 Hague Declaration</td>
<td>legally binding</td>
<td>banned use</td>
<td>failed</td>
<td>civilian casualties</td>
<td></td>
</tr>
<tr>
<td>World War I to World War II</td>
<td>sawback bayonets</td>
<td>World War I</td>
<td>tacit cooperation on the battlefield</td>
<td>no explicit agreement</td>
<td>norm against possession</td>
<td>successful</td>
<td>unnecessary suffering</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------------------------------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>chemical and bacteriological weapons</td>
<td>1925</td>
<td>1925 Geneva Gas and Bacteriological Protocol</td>
<td>legally binding</td>
<td>banned use</td>
<td>largely successful in restraining battlefield use in WWII</td>
<td>unnecessary suffering</td>
<td></td>
</tr>
<tr>
<td>submarines</td>
<td>1899; 1921-1922</td>
<td>1899 Hague convention; 1921-1922 Washington Naval Conference</td>
<td>never ratified</td>
<td>attempted bans - never ratified</td>
<td>failed - treaty never ratified</td>
<td>civilian casualties</td>
<td></td>
</tr>
<tr>
<td>submarines</td>
<td>1907; 1930; 1936</td>
<td>1907 Hague Declaration; 1930 London Naval Treaty; 1936 London Protocol</td>
<td>legally binding</td>
<td>regulated use</td>
<td>failed - compliance collapsed in war</td>
<td>civilian casualties</td>
<td></td>
</tr>
<tr>
<td>size of navies</td>
<td>1922; 1930; 1936</td>
<td>1922 Washington Naval Treaty; 1930 London Naval Treaty; 1936 Second London Naval Treaty</td>
<td>legally binding</td>
<td>limited quantities and size of ships</td>
<td>short-lived</td>
<td>limit arms races</td>
<td></td>
</tr>
<tr>
<td>nuclear tests</td>
<td>1963; 1967; 1985; 1995; 1996</td>
<td>Limited Test Ban Treaty; Treaty of Tlatelolco; Treaty of Bangkok; Treaty of Rarotonga; Treaty of Treaty of</td>
<td>legally binding</td>
<td>restricted testing</td>
<td>generally successful, with some exceptions</td>
<td>effects on civilians; limit arms races</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>Year(s)</td>
<td>Treaty/Agreement</td>
<td>Binding Status</td>
<td>Limitations</td>
<td>Outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------</td>
<td>----------------------------------------------------</td>
<td>-----------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weapons in Antarctica</td>
<td>1959</td>
<td>Antarctic Treaty</td>
<td>Legally binding</td>
<td>Banned deployment</td>
<td>Successful limit arms races</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weapons of mass destruction in space</td>
<td>1967</td>
<td>Outer Space Treaty</td>
<td>Legally binding</td>
<td>Banned deployment</td>
<td>Successful strategic stability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weapons on the moon</td>
<td>1967</td>
<td>Outer Space Treaty</td>
<td>Legally binding</td>
<td>Banned deployment</td>
<td>Successful limit arms races</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear-free zones</td>
<td>1967;</td>
<td>Treaty of Tlatelolco;</td>
<td>Legally binding</td>
<td>Banned developing, manufacturing, possessing, or stationing</td>
<td>Successful limit arms races</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1985;</td>
<td>Treaty of Rarotonga;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1995;</td>
<td>Treaty of Bangkok;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>Treaty of Pelindaba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear weapons</td>
<td>1970</td>
<td>Nuclear Non-Proliferation Treaty</td>
<td>Legally binding</td>
<td>Banned proliferation</td>
<td>Generally successful, with some exceptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear weapons on the seabed</td>
<td>1971</td>
<td>Seabed Treaty</td>
<td>Legally binding</td>
<td>Banned deployment</td>
<td>Successful strategic stability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ballistic missile defenses</td>
<td>1972</td>
<td>Anti-ballistic Missile Treaty</td>
<td>Legally binding</td>
<td>Limited deployment</td>
<td>Successful during Cold War; collapsed in multipolar world</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

313
<table>
<thead>
<tr>
<th>Issue</th>
<th>Year</th>
<th>Agreement</th>
<th>Nature</th>
<th>Outcome</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological weapons</td>
<td>1972</td>
<td>Biological Weapons Convention</td>
<td>legally binding</td>
<td>banned development, production, stockpiling, and use</td>
<td>generally successful, with some exceptions</td>
</tr>
<tr>
<td>Using the environment as a weapon</td>
<td>1976</td>
<td>Environmental Modification Convention</td>
<td>legally binding</td>
<td>banned use</td>
<td>successful</td>
</tr>
<tr>
<td>Anti-satellite weapons</td>
<td>1970s &amp; 1980s</td>
<td>tacit cooperation between U.S. and U.S.S.R.</td>
<td>no explicit agreement</td>
<td>norm against deployment</td>
<td>successful, but currently threatened in multipolar world</td>
</tr>
<tr>
<td>Neutron bombs</td>
<td>1970s</td>
<td>tacit cooperation between U.S. and U.S.S.R.</td>
<td>no explicit agreement</td>
<td>norm against deployment</td>
<td>successful</td>
</tr>
<tr>
<td>Land mines</td>
<td>1980</td>
<td>CCW Protocol II</td>
<td>legally binding</td>
<td>regulated use</td>
<td>unsuccessful</td>
</tr>
<tr>
<td>Incendiary weapons</td>
<td>1980</td>
<td>CCW Protocol III</td>
<td>legally binding</td>
<td>regulated use</td>
<td>mixed success</td>
</tr>
<tr>
<td>Chemical and biological weapons</td>
<td>1985</td>
<td>Australia Group</td>
<td>not legally binding</td>
<td>banned proliferation</td>
<td>generally successful, with some exceptions</td>
</tr>
<tr>
<td>Ballistic and cruise missiles</td>
<td>1987; 2002</td>
<td>Missile Technology Control Regime; Hague</td>
<td>not legally binding</td>
<td>limited proliferation</td>
<td>has had some success</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>nuclear weapons and launcher quantities</td>
<td>1972; 1979; 1991; 2002; 2011</td>
<td>SALT I; SALT II; START; SORT; New START</td>
<td>legally limited quantities</td>
<td>SALT I; SALT II; START; SORT; New START</td>
<td>successful limit arms races</td>
</tr>
<tr>
<td>conventional air and ground forces</td>
<td>1991</td>
<td>Conventional Forces in Europe</td>
<td>legally limited quantities</td>
<td>Conventional Forces in Europe</td>
<td>collapsed in multipolar world</td>
</tr>
<tr>
<td>chemical weapons</td>
<td>1993</td>
<td>Chemical Weapons Convention</td>
<td>legally banned development, production, stockpiling, and use</td>
<td>Chemical Weapons Convention</td>
<td>generally successful, with some exceptions</td>
</tr>
<tr>
<td>blinding lasers</td>
<td>1995</td>
<td>CCW Protocol IV</td>
<td>legally banned use</td>
<td>CCW Protocol IV</td>
<td>successful</td>
</tr>
<tr>
<td>conventional weapons</td>
<td>1996</td>
<td>Wassenaar Arrangement</td>
<td>not legally limited proliferation</td>
<td>Wassenaar Arrangement</td>
<td>has had some success</td>
</tr>
<tr>
<td>land mines</td>
<td>1997</td>
<td>Mine Ban Treaty (Ottawa Treaty)</td>
<td>legally banned development, production, stockpiling, and use</td>
<td>Mine Ban Treaty (Ottawa Treaty)</td>
<td>generally successful, with some exceptions</td>
</tr>
<tr>
<td>cluster munitions</td>
<td>2008</td>
<td>Convention on Cluster Munitions</td>
<td>legally banned development, production, stockpiling, and use</td>
<td>Convention on Cluster Munitions</td>
<td>generally successful, with some exceptions</td>
</tr>
</tbody>
</table>
Interviews


Clune, Jeff. Associate Professor of Computer Science, University of Wyoming. Interview by author. Telephone. September 28, 2016.

Cummings, Mary “Missy”. Professor of Electrical and Computer Engineering, Duke University. Interview by author. Durham, NC. June 1, 2016.


Kennedy, William. Associate Professor in the Center for Social Complexity and the Department of Computational and Data Science, George Mason University. Interview by author. Fairfax, VA. December 8, 2015


Walker, Mike. DARPA program manager. Interview by author. Telephone. December 5, 2016.

Original Sources

10 U.S.C. 2431 Sec. 224.


“Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques.” Bureau of International Security and
https://www.state.gov/t/isn/4783.htm#treaty.


DARPA. “Collaborative Operations in Denied Environment (CODE): Phase 2 Concept Video.” https://www.youtube.com/watch?v=BPBuE6fMBnE.

DARPA. “Collaborative Operations in Denied Environment (CODE): Test of Phase 1 Human-System Interface.” https://www.youtube.com/watch?v=o8AFuiO6ZSs&feature=youtu.be.


Department of Justice. “Futures Trader Pleads Guilty to Illegally Manipulating the Futures Market in Connection With 2010 ‘Flash Crash.’” November 9, 2016.
Department of State Telegram Transmitting Letter From Chairman Khrushchev to President Kennedy. October 26, 1962.
http://microsites.jfklibrary.org/cmc/oct26/doc4.html


Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.”


Headquarters, Department of the Army. Field Manual 100-5 (June 1993).

ICRC. “Customary International Humanitarian Law Rules, IV. Weapons.”
https://www.icrc.org/customary-ihl/eng/docs/v1_rul.


Israel Aerospace Industries. “Harpy NG.”


http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Kataev.PDF.


http://nsarchive.gwu.edu/nukevault/ebb285/vol%20II%20Korobushin.PDF.

Laws of Manu, Chapter VII: 90. Translated by G. Buhler.
http://sourcebooks.fordham.edu/halsall/india/manu-full.asp.


Ministry of Defence. “Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710.”


"Treaty Between The United States of America and The Union of Soviet Socialist Republics on the Limitation of Strategic Offensive Arms (SALT II)." Bureau of


Secondary Sources


Bostrom, Nick. “Ethical Issues in Advanced Artificial Intelligence.”


https://forallsecure.com/blog/2016/08/06/mayhem-wins-darpa-cgc/.


http://eprints.lse.ac.uk/87629/1/Coker_Human%20Thing.pdf.

Colby, Elbridge A. “Defining Strategic Stability: Reconciling Stability and Deterrence.”


“Combat tests in Syria brought to light deficiencies of Russian unmanned mini-tank.”


Cox, Jeff. “‘Knight-Mare’: Trading Glitches May Just Get Worse.” August 2, 2012. 


Ganguly, Sumit and S. Paul Kapur, Eds. *Nuclear Proliferation In South Asia: Crisis Behaviour and the Bomb* (Routledge, 2009).


http://germanstories.vcu.edu/goethe/zauber_dual.html.

https://www.youtube.com/watch?v=Pq4A2mPCBoY.


https://web.archive.org/web/20121120161047/http://www.airforce-


April 9, 2013.
http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Sessio

Heyns, Christof. “Robot wars: after drones, a line we must not cross.”
http://www.theguardian.com/commentisfree/2013/jun/19/robot-wars-drones-
life-death-decisions.

Hoffman, David. “‘I Had a Funny Feeling in My Gut.’” Washington Post. February 10,
srv/inatl/longterm/coldwar/shatter021099b.htm.

Horowitz, Michael C. “Artificial Intelligence, International Competition, and the
https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-
the-balance-of-power/.

Horowitz, Michael C. “When speed kills: Lethal autonomous weapons systems,
[forthcoming].

Horowitz, Michael and Paul Scharre. “Meaningful Human Control in Weapon Systems:
A Primer.” Center for a New American Security. Washington, DC. March 16,
weapon-systems-a-primer.

Horton, Alex. “Questions Hover over Army Drone’s 630-Mile Odyssey across Western
US.” Stars and Stripes. March 1, 2017.
https://www.stripes.com/news/questions-hover-over-army-drone-s-630-mile-
odyssey-across-western-us-1.456505#.WLby0hLyyv5Z.

robots-lost-high-frequency-tradings-rise-and-fall.


Kenney, Bin. “Jeopardy!—The IBM Challenge (Day 1—February 14).” Video. https://www.youtube.com/watch?v=i-vMW_Ce51w.


https://www.youtube.com/watch?v=ZBjmm48zwQU.


*Environmental Science and Technology* 47 (2013).

Lockie, Alex. “Russia Claims Its Deadly T-14 Armata Tank Is in Full Production.”

“Lost submarines of the Soviet Union.” Wikipedia.

“M898 SADARM (Sense and Destroy Armor).” GlobalSecurity.org.


McNamara, Robert. Interview included as special feature on Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (DVD). Columbia Tristar Home Entertainment. 2004 [1964].


Murphy, Tom, VII. “The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel . . . after that it gets a little tricky.” https://www.cs.cmu.edu/~tom7/mario/mario.pdf


“TV (G7es) Acoustic Homing Torpedo.” German U-Boat.

http://money.cnn.com/2010/05/06/markets/markets_newyork/.


