Autonomous Weapons and Operational Risk

Ethical Autonomy Project
February 2016

Paul Scharre
Senior Fellow and Director of the 20YY Future of Warfare Initiative
Center for a New American Security

Center for a New American Security

About the author

Paul Scharre is a senior fellow and director of the 20YY Future of Warfare Initiative at CNAS.

Acknowledgements

I would like to thank William Kennedy of George Mason University, John Hawley of the Army Research Laboratory, and John Borrie of the United Nations Institute for Disarmament Research for their helpful feedback on this report. A special thanks to CNAS' Loren Schulman, Maura McCarthy, and Melody Cook for their work in the development, production, and design of this report. This research is made possible by the generous support of the John D. and Catherine T. MacArthur Foundation. Any errors of fact or omission are mine alone. CNAS does not take institutional positions.

Also in this series

"An Introduction to Autonomy in Weapon Systems," by Paul Scharre and Michael C. Horowitz (February 2015)

"Meaningful Human Control in Weapon Systems: A Primer," by Michael C. Horowitz and Paul Scharre (March 2015)

"Autonomous Weapons at the UN: A Primer for Delegates," by Paul Scharre, Michael C. Horowitz, and Kelley Sayler (April 2015)

Table of Contents

I. Preface

Automation is increasing in a wide range of military systems, including future weapons. In response, a growing number of voices are calling for urgent discussion on the appropriate role of human and machine decision-making in the use of lethal force. The Center for a New American Security's Ethical Autonomy Project examines the legal, moral, ethical, and policy issues associated with autonomous weapons—weapons that would select and engage targets on their own. Previous papers in this series have examined technical aspects of autonomous weapons and the concept of "meaningful human control."

This paper is aimed at helping defense professionals think clearly and objectively about possible risks associated with autonomous weapons. Autonomous weapons generally do not exist, and their military costs and benefits can be speculated but are not yet clearly known. What is clear, however, is that they raise novel questions of risk. The essence of autonomy is delegating a task previously done by a person to a machine. This raises the important question of how to retain effective human control over the machine's behavior and the risks—both the probability and consequences—associated with a loss of control.

In doing so, this paper examines the risks of autonomous weapons *relative to semi-autonomous weapons* that would retain a human "in the loop" for selecting and engaging specific targets. Autonomous weapons cannot be viewed in a vacuum. War is dangerous, and weapons that are intended to be deadly to an opponent often can be quite dangerous to the user or friendly forces as well. This paper therefore aims to answer the question: How is the risk of using autonomous weapons *different* from other weapons?

This paper joins an extensive and burgeoning literature on autonomous weapons, much of which addresses legal, moral, or ethical considerations. While questions of risk and control may have legal or other implications, this paper does not address these, instead focusing solely on safety. For example, "meaningful human control" is one paradigm that has been presented for thinking about human control and autonomous weapons. This paper does not attempt to address the very important issues of legal and moral responsibility and accountability associated with what makes human control "meaningful." Rather, it focuses solely on the safety and reliability questions of what makes human control *effective*.[1]

In focusing on safety and risk, this paper assumes that the autonomous system in question is capable of performing the basic functions required of it under most operating conditions. In the case of autonomous weapons, this means completing engagements in a manner consistent with the laws of war (international humanitarian law). While this is a contestable assumption and may depend to a large degree on the operating environment, target set, and timeframe for autonomy being

[1] For a related analysis of safety aspects of autonomous weapons, see John Borrie, "Safety aspects of 'meaningful human control': Catastrophic accidents in complex systems," UNIDIR Conference: Weapons, Technology and Human Control, New York, October 16, 2014.

envisioned, it is clear that there are at least limited situations where this would be true today. In situations where civilians were not present, such as undersea, and when targeting military vehicles, autonomous weapons could likely be used lawfully in limited situations today, provided their operation was sufficiently restricted in time and space.

This paper does not examine *how* autonomous weapons might be able to complete engagements lawfully or under what conditions this might be possible in the future. Rather, it simply assumes that lawful employment of autonomous weapons is feasible, at least for isolated situations. This paper then examines the *risk* in employing autonomous weapons in these situations, relative to semi-autonomous weapons that would retain a human in the loop, and provides guidance for militaries and policymakers in evaluating these risks. As militaries incorporate increasing autonomy into future weapon systems, a clear-eyed understanding of the associated operational risks is vital. Additionally, risk and safety is a critical dimension to autonomous weapons that should be incorporated into ongoing international discussions at the United Nations and other forums.

II. Executive Summary

Autonomous weapon systems—potential future weapons that would select and engage targets on their own—raise a host of legal, ethical, and moral questions. They also raise critically important considerations regarding safety and risk.

Autonomous weapons have a qualitatively different degree of risk than equivalent semi-autonomous weapons that would retain a human in the loop. The consequences of a failure that causes the weapon to engage an inappropriate target could be far greater with an autonomous weapon. The result could be fratricide, civilian casualties, or unintended escalation in a crisis.

Humans are not immune from errors, and semi-autonomous weapons can also fail. However for semi-autonomous weapons, requiring a human in the loop to authorize each engagement creates a natural fail-safe. If the weapon system begins to fail, the human controller can modify the weapon system's operation or halt the engagement before further damage is done.

With an autonomous weapon, however, the damage potential before a human controller is able to intervene could be far greater. In the most extreme case, an autonomous weapon could continue engaging inappropriate targets until it exhausts its magazine, potentially over a wide area. If the failure mode is replicated in other autonomous weapons of the same type, a military could face the disturbing prospect of large numbers of autonomous weapons failing simultaneously, with potentially catastrophic consequences.

From an operational standpoint, autonomous weapons pose a novel risk of *mass fratricide*, with large numbers of weapons turning on friendly forces. This could be because of hacking, enemy behavioral manipulation, unexpected interactions with the environment, or simple malfunctions or software errors. Moreover, as the complexity of the system increases, it becomes increasingly difficult to verify the system's behavior under all possible conditions; the number of potential interactions within the system and with its environment is simply too large.

While these risks can be mitigated to some extent through better system design, software verification and validation, test and evaluation, and user training, these risks cannot be eliminated entirely. Complex tightly coupled systems are inherently vulnerable to "normal accidents." The risk of accidents can be reduced, but never can be entirely eliminated.

Militaries considering autonomous weapons must carefully weigh these risks against military utility and the potential disadvantage of keeping a human in the loop as a fail-safe, if nothing else. Human decision-making and automation are not mutually exclusive, however. "Centaur" human-machine teaming cognitive architectures can leverage the predictability, reliability, and speed of automation while retaining the robustness and flexibility of human intelligence. Whenever possible, human-machine teaming will be preferred.

III. Introduction: Robutopia or Robocalypse?

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).[2] 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.[3]

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. While these fears are fed by dystopian science fiction tales (after all, a utopian future is a boring story), these concerns also are rooted in our everyday experience with automated systems. 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. Unlike humans, autonomous systems lack the ability to step outside their instructions and employ "common sense," adapting to the situation at hand.[4]

These two intuitions regarding autonomous systems implicitly shape much of the discourse on autonomous weapons, with some viewing them as potentially beneficial technologies that could make war more precise and humane, and others viewing them as dangerous technologies that could lead to catastrophe. Which view is correct? Will autonomous weapons lead to a robutopia or robocalypse?

Both of these conflicting intuitions have a basis in reality, and in many situations autonomous systems will demonstrate both features. Provided they can adequately perform the task, under normal operating conditions they may very well perform better than humans. However, their brittle nature means that if pushed beyond the bounds of their programming, they may fail, and fail badly.

---

[2] "Accidents or Unintentional Injuries," Center for Disease Control and Prevention, http://www.cdc.gov/nchs/fastats/accidental-injury.htm.

[3] For example "Intelligent Drive," Mercedes-Benz, https://www.mbusa.com/mercedes/technology/videos/detail/title-safety/videoId-fc0835ab8d127410VgnVCM100000ccec1e35RCRD.

[4] In theory, future artificial general intelligence systems could consider the broader context and adapt to novel situations. However, such systems, if they could be built, would introduce other potentially more serious challenges. For more on the risks associated with artificial general intelligence, see Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

Autonomous systems lack the flexibility humans have to adapt to novel circumstances. This may mean that in unexpected situations, autonomous systems make mistakes that humans would not have.

As militaries think through how they incorporate autonomy into future weapons, a clear-eyed assessment of the risks associated with autonomous weapons is required. Accidents with autonomous weapons could lead to civilian casualties, fratricide, or unintended escalation in a crisis. Understanding these risks is important for policymakers and acquisition professionals weighing whether to build autonomous weapons, as well as for military commanders who are responsible for the weapons they deploy on the battlefield.

Of course, humans are far from perfect in war. They make mistakes that lead to fratricide and civilian casualties as well, and humans also commit deliberate acts of atrocity.[5] There are important qualitative differences between the risks associated with semi-autonomous (human in the loop) weapons and autonomous weapons, however:

- Without a human in the loop to act as a fail-safe, the consequences of failure with an autonomous weapon could be far more severe than an equivalent semi-autonomous weapon.

- For extremely simple autonomous weapons, these risks may be manageable if human operators employ them only in limited, controlled contexts. As autonomous weapons increase in complexity, however, it may be more difficult for human operators to fully understand the boundaries of their behavior and accurately predict under what conditions failures might occur, even if they are unlikely.

- While improved test and evaluation can mitigate these risks somewhat, they cannot be eliminated entirely. Failures in complex systems may not be likely, but over a long enough time horizon they are inevitable.

[5] Ronald C. Arkin, "Warfighting Robots Could Reduce Civilian Casualties, so Calling for a Ban Now Is Premature," IEEE Spectrum, August 5, 2015, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/autonomous-robotic-weapons-could-reduce-civilian-casualties.

IV. Controlling Autonomous Systems

In order to better understand the risks associated with autonomous weapons, we can first examine the nature of control over autonomous systems in general. An autonomous system is one that, once activated, performs a task on its own. Everyday examples range from simple systems like toasters and thermostats to more sophisticated systems like automobile intelligent cruise control or airplane autopilots. The *risk* in employing an autonomous system is that the system might not perform the task in a manner that the human operator intended.

There are a number of reasons why an autonomous system might begin performing inappropriately, from simple malfunctions and software bugs to more complex system failures, changing environmental conditions, hacking, and human error. When these failures can be anticipated in advance, human operators can account for these limitations. When failures are unanticipated, however, the result can be autonomous systems that slip out of control. Understanding the likelihood and consequences of a loss of control is essential to assessing the risk in employing autonomous systems.

Maintaining effective control over autonomous systems

Human operators will want to ensure that autonomous systems perform in a way consistent with their intentions. Because autonomous systems in many cases do not have real-time human supervision, maintaining *effective control* over the system has two components:

1. The ability of the human operator to accurately predict the autonomous system's behavior in the environment in which it is being used. This includes its limitations and the conditions under which it will fail. This allows the human operator to employ the autonomous system only in situations where it will perform appropriately.

2. The ability of the human operator to undertake corrective action if/when the autonomous system fails to behave in accordance with the human operator's intentions.

A *failure* with an autonomous system is a loss of effective control—a situation in which the autonomous system no longer is behaving in accordance with human operator intention.

Risk, in this context, refers to the risk of failure, both the probability and consequences of a loss of control:

- The *probability of failure* is the likelihood of the system behaving in a manner inconsistent with human operator intentions in a particular environment.

- The *consequence of failure* is the potential damage the autonomous system could do in that environment until such time as the human operator can undertake corrective action to bring

the system back in line with human operator intentions or the autonomous system ceases operation.

Autonomous systems can vary in the type of task they perform, their level of complexity, and degree of the human operator's interaction with the system. As these aspects of the system change, the risks of employing an autonomous system change as well.

The inherent hazard of a system depends on the task being performed

Autonomous systems can perform a wide variety of tasks, from driving cars to regulating temperature or making toast. The inherent hazard of a system is the potential consequence if the autonomous system performs that task incorrectly. This depends on both the task being performed and its operating environment. The consequences of a failure with an autonomous car are far more potentially severe than a toaster failing to properly cook bread. The environment in which the system is operating is also a key component of the inherent hazard of the system. The hazard associated with an autonomous car driving on a closed-circuit track is much less severe than one driving through crowded city streets with pedestrians.

The time between failure and corrective action depends on the type of human control

The type of human control over the system is a key variable affecting the potential consequences of a failure with an autonomous system. There are three broad types of control humans can exercise:[6]

- **Semi-autonomous operation**, where the machine performs a task and then stops and waits for approval from the human operator before continuing. This control type is often referred to as "human in the loop."

- **Supervised autonomous operation**, where the machine, once activated, performs a task under the supervision of a human and will continue performing the task unless the human operator intervenes to halt its operation. This control type is often referred to as "human on the loop."

- **Fully autonomous operation**, where the machine, once activated, performs a task and the human operator does not have the ability to supervise its operation and intervene in the event of system failure. This control type is often referred to as "human out of the loop."

The control type affects the human operator's ability to undertake corrective action if the system fails.

[6] For more on this topic, see Paul Scharre and Michael C. Horowitz, "An Introduction to Autonomy in Weapon Systems," (Center for a New American Security, February 2015).

In a semi-autonomous system, after each task the machine stops and waits for human approval before continuing. Under this control type, the human operator has the ability to observe the machine's actions in the environment and confirm that the behavior is appropriate before continuing.[7]

In supervised autonomous operation, in principle the human operator has the ability to intervene if necessary. In practice, there is likely to be some time delay between when a failure occurs and when the human is able to actually correct the behavior of the system. This could occur because of a time delay in communications or because it may take the human some period of time to understand that the system is performing inappropriately and determine the appropriate corrective action.

For fully autonomous systems, the human operator lacks the ability to observe the autonomous system's behavior and undertake corrective action in sufficient time if the system fails to perform appropriately. Presumably at some point in time the human operator will become aware of how the system performed. For example, a household thermostat is operating "fully autonomously" while one is away from the home. Once one returns home, one discovers whether the thermostat was performing as one intended or not.[8]

Thus a key element of risk in autonomous systems is the time between when a system begins failing (performing in a manner other than what the human operator intended) and when the human operator can undertake corrective action. Even in fully autonomous systems, presumably the system ceases operation at a certain point in time once the task is complete and the results of its actions can be observed.

The damage potential of a system depends on its inherent hazard and the time from failure to corrective action

When it comes to risk, we are concerned with an autonomous system's *damage potential.* Damage potential is the amount of damage an autonomous system could do, if it failed to perform appropriately, before a human operator could take corrective action. Damage potential depends upon the inherent hazard of the system—the type of task being performed and the environment in which it is operating—as well as the control type. For supervised autonomous systems, the speed of the system's operation and any potential time delays also are significant factors. A system that in principle has a human on the loop to intervene in the event of system failure might in practice still have a high damage potential if the system performs tasks much more rapidly than human operators can react.

Consider, for example, an autonomous car. The damage potential of an autonomous car can vary significantly depending on these variables. Not only does a car driving through crowded city streets

[7] This assumes that the human operator is given sufficient information about the system's behavior to make an informed decision, not merely pro forma approval.

[8] By contrast, a web-enabled thermostat that one could monitor from a smartphone or other device would be operating in a human-supervised autonomous mode when it was being observed, provided that one could change the thermostat remotely.

have a higher inherent hazard than one on a closed track, the type of human control over the car could significantly change the damage potential. A self-driving car that is equipped with a steering wheel and brake to allow the human operator to take control and stop the vehicle (human-supervised autonomy) has, in principle, lower damage potential than a fully autonomous car where the human is merely a passenger along for the ride.

The speed of interactions matters significantly, however. Giving the human operator the ability to grab the wheel of an autonomous vehicle traveling at highway speeds in dense traffic, particularly if the operator is not paying attention, is merely the illusion of control. Conversely, a brake and steering wheel on an autonomous car moving slowly under the supervision of an attentive human operator might add real value by allowing the human to function as an additional fail-safe. The driver may not be able to prevent all accidents (after all, humans are not great drivers even when directly in control of the vehicle), but he or she could prevent an autonomous car from running rampant, senselessly mowing down pedestrians.

An unfortunate reality of both supervised autonomous and even semi-autonomous operation is that the human operator may not become aware that the system is failing until after a failure occurs. A human in the loop or on the loop will not necessarily prevent failures from occurring. However, the ability of a human to undertake corrective action can help limit the damage potential of a system if it fails. Thus, in these circumstances, the human functions as a fail-safe. The human operator cannot necessarily prevent failure, but he or she can help ensure that if or when the system fails, the damage is limited.

Increasing complexity can complicate the human operator's ability to accurately predict system behavior

Another critically important dimension of autonomous systems is their degree of complexity—both the complexity of the system itself and the complexity of the environment in which it is operating. Complexity matters because it affects the human operator's ability to predict the behavior of the system.

In general, simpler systems operating in simpler environments will be easier to predict. Simpler systems are likely to be more limited in the types of operations they can perform, and their operation is likely to be more transparent to trained operators. However, the range of environments in which they can operate is also likely to be more limited. To operate in a broader range of environments or accomplish more difficult tasks, more sophisticated autonomous systems are needed, but by necessity they are more complex. This complexity can make the system less transparent even to well-trained operators. As a result, predicting the system's behavior, particularly when operating in complex and unstructured real-world environments, can be more challenging.

In ordinary speech, we often use words like "automatic," "automated," "autonomous," and "intelligent" to refer to a spectrum of complexity and sophistication in autonomous systems.

We tend to use words like **automatic** to refer to simple, threshold-based systems with easily predictable reflexive responses to external input. Examples include toasters, mines, or old mechanical thermostats.

- We tend to refer to more complex, rule-based systems as **automated.** Examples include self-driving vehicles, modern computers, many modern military weapon systems, and programmable thermostats.

- The term **autonomous** is sometimes reserved for systems that exhibit some degree of learning, adaptation, or evolutionary behavior. Others, however, might use the term "autonomous" to refer to complex rule-based systems that exhibit goal-oriented behavior, systems that some might call "automated." Examples of learning systems include robots that teach themselves how to move around their environment or the Nest "learning thermostat."[9]

- Finally, we sometimes refer to autonomous systems that are capable of human-level cognitive tasks, at least for narrow problems, as "**intelligent**." Examples include IBM's chess-playing computer Deep Blue and IBM's *Jeopardy!* computer contestant Watson.[10]

There are no clear boundaries between these categories, and people can disagree on what to call any given system. Further complicating matters, the degree of complexity of a system is independent of its human-machine control type (semi-, supervised, or fully autonomous) and independent of the task it performs. This can lead to significant semantic confusion, as a system may be "fully autonomous" in the sense of operating without human supervision, but its functionality may be quite simple, leading some to refer to it as "automatic" or "automated."

From a risk perspective, increasing sophistication and complexity has both benefits and drawbacks. On the one hand, more sophisticated and intelligent autonomous systems are desirable, since they are generally more capable of performing a broader array of tasks under a wider set of conditions. If a main limitation of autonomous systems is their brittleness—their inability to adapt to unexpected conditions and step outside their programming—more intelligent systems are one potential solution. More sophisticated autonomous systems can account for a wider range of variables, allowing them to expand the number and complexity of situations in which they can operate effectively. More complex systems can still fail if pushed outside the bounds of their intended operating environment, but the scope of situations they can handle is increased.

While more capable autonomous systems are desirable, their additional complexity is a double-edged sword, however. As a system becomes more complex, it becomes increasingly difficult for a human operator to predict precisely what the autonomous system might do in any given situation. This is the case even for well-trained operators. In many ways, the intent of delegating control to

[9] "Nest Thermostat," Nest, https://nest.com/thermostat/meet-nest-thermostat/
[10] "What is Watson?," IBM, http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html.

autonomous systems is to allow them the flexibility to respond to specific, unforeseen situations in the best manner possible in order to accomplish a goal. Building a car that can navigate through crowded city streets is difficult precisely because the human operator cannot predict in advance exactly when the car must stop, go, turn left, or turn right. Rather, the system itself must be sophisticated enough to be able to make those determinations on its own according to the environmental conditions it encounters. This means that the human operator must trust that the autonomous system will execute its tasks, if not perfectly, at least in an acceptable manner.

It is clear how systems that exhibit learning, evolutionary, or emergent behavior could give rise to surprising behaviors.[11] However, even sufficiently complex rule-based (automated) systems can act in unexpected ways. This could occur for a variety of reasons:

- **Malfunctions and bugs:** As a system becomes more complex, the sheer number of mechanical parts and lines of code grows, increasing the number of elements that could malfunction or be coded improperly.

    - Studies have evaluated the software industry average error rate at 15-50 errors per 1,000 lines of code. Rigorous internal test and evaluation has been able to reduce the error rate to 0.1-0.5 errors per 1,000 lines of code in some cases.[12] However, in systems with millions of lines of code, some errors are inevitable.[13]

- **System failures:** System failures occur not from the breakdown of any one given part, but from unanticipated interactions between elements of a system. Verifying all possible combinations of the internal workings of the system becomes increasingly difficult as the system's complexity increases.

    - A recent report on autonomy by the U.S. Air Force Office of the Chief Scientist highlighted the need for new techniques for the verification and validation of autonomous software as a "critical" issue for the Air Force. "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."[14]

---

[11] Emergent behavior could come from individual systems or from groups or swarms of simpler systems coordinating their actions together, similar to ants, termites, or bees. For more on military applications of swarming, see Paul Scharre, "Robotics on the Battlefield – Part II: The Coming Swarm," Center for a New American Security, October 2014, http://www.cnas.org/sites/default/files/publications-pdf/CNAS_TheComingSwarm_Scharre.pdf.

[12] Steve McConnell, *Code Complete: A Practical Handbook of Software Construction* (Redmond, WA: Microsoft Press, 2004) http://www.amazon.com/Code-Complete-Practical-Handbook-Construction/dp/0735619670.

[13] The space shuttle is an interesting exception that proves the rule. NASA has been able to drive the number of errors on space shuttle code down to zero through a labor-intensive process employing teams of engineers. However, the space shuttle has only approximately 500,000 lines of code, and this process would be entirely unfeasible for more complex systems using millions of lines of code. The F-35 Joint Strike Fighter, for example, has over 20 million lines of code. Charles Fishman, "They Write the Right Stuff," FastCompany.com, December 31, 1996, http://www.fastcompany.com/28121/they-write-right-stuff.

[14] U.S. Air Force Office of the Chief Scientist, *Autonomous Horizons: System Autonomy in the Air Force – A Path to the Future* (June 2015), 23, http://www.af.mil/Portals/1/documents/SECAF/AutonomousHorizons.pdf?timestamp=1435068339702.

- **Systems are not transparent to human operators:** While one of the benefits of automation in many cases is reducing the potential for human error, a negative side effect to greater complexity is that the functioning of the system may be increasingly opaque to even trained users.

    - On June 1, 2009, Air France Flight 447 from Rio de Janeiro to Paris crashed into the Atlantic Ocean, killing all 228 people on board. Four hours into the flight, air speed indicators on the wings froze in bad weather, a rare but non-serious instrumentation error that disengaged the autopilot and put control back in the hands of the pilots. Eleven seconds following the autopilot disengagement, the pilots correctly identified that they had lost the airspeed indicators. At this point in time, the aircraft was flying normally, at appropriate speeds and full altitude, and there was no emergency. Inexplicably, however, the pilots began a series of errors that resulted in a stall, causing the aircraft to crash into the ocean. Throughout the incident, the pilots continually misinterpreted data from the airplane and misunderstood the aircraft's behavior. In part, poor user interfaces and opaque automated processes on the aircraft, even while flown manually, contributed to this lack of understanding. The complexity of the aircraft created problems of transparency that likely would not have existed in a similar situation on a simpler aircraft.[15]

- **Unanticipated interactions with the environment**: As the complexity of the system and/or its operating environment increases, the number of potential interactions increases dramatically. This can make testing the autonomous system's operation under every possible environmental condition effectively impossible.

    - On their first deployment to the Pacific, eight F-22 fighter jets experienced a Y2K-like total computer failure when crossing the international dateline. All onboard computer systems shut down, and the result was nearly a catastrophic loss of the aircraft. While the existence of the international dateline could clearly be anticipated, the interaction of the dateline with the software was not identified in testing.[16]

- **Adversarial hacking**: In an adversarial environment, such as in war, enemies will likely attempt to exploit vulnerabilities of the system, whether through hacking, spoofing (sending false data), or behavioral hacking (taking advantage of predictable behaviors to "trick" the system into performing a certain way). While any computer system is, in principle, susceptible to hacking, greater complexity can make it harder to identify and correct any vulnerabilities.

---

[15] The full, official accident report by French authorities is, "Final Report: On the accidents of 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight 447 Rio de Janeiro – Paris," Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, [English translation], 2012, http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf. For a shorter and more readable summary of events, read William Langewiesche, "The Human Factor," *Vanity Fair,* October 2014, http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash; and Nick Ross and Neil Tweedie, "Air France Flight 447: 'Damn it, We're Going to Crash,'" *The Telegraph*, April 28, 2012, http://www.telegraph.co.uk/technology/9231855/Air-France-Flight-447-Damn-it-were-going-to-crash.html.

[16] Remarks by Air Force retired Major General Don Sheppard on "This Week at War," *CNN,* February 24, 2007, http://transcripts.cnn.com/TRANSCRIPTS/0702/24/tww.01.html.

- In 2015, two hackers revealed that they had discovered software vulnerabilities that allowed them to remotely hack certain automobiles while they were on the road. This allowed the hackers to remotely take control of critical driving functions including the transmission, steering, and brakes.[17]

Neural networks and the black box

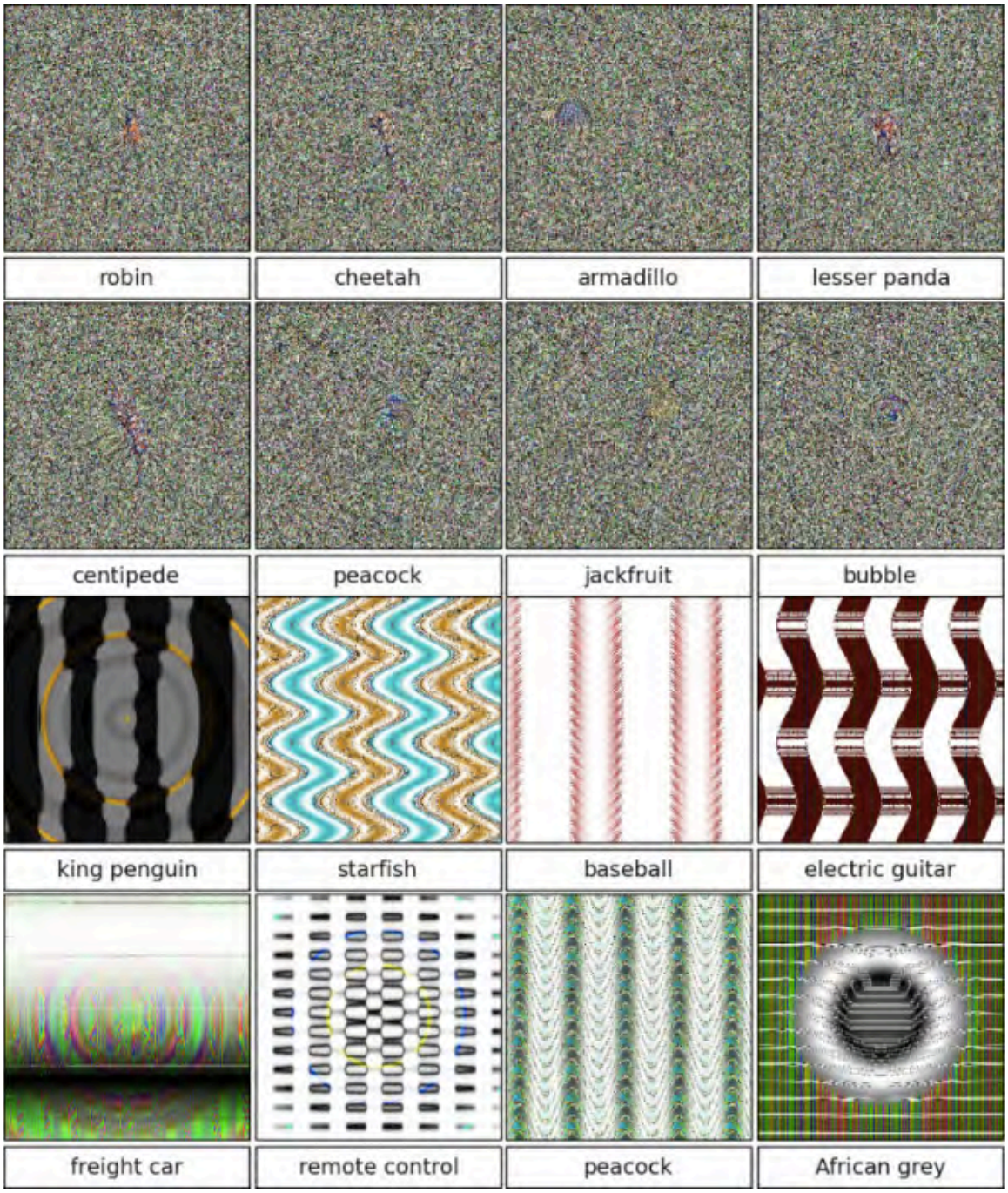
The challenge of complexity, while problematic for complex rule-based systems, is even more difficult for cutting-edge artificial intelligence (AI) systems that employ neural networks. Neural networks do not perform rule-based calculations like most computers. Instead, they learn by exposure to large data sets. As a result, the internal structure of the network that generates output can be opaque to the designers—a "black box." Even more unsettling, for reasons that may not be entirely clear to AI researchers, the neural network sometimes can yield odd, counterintuitive results.

A study of visual classification AIs using neural networks found that while the AIs were able to generally identify objects as well as humans, in some cases the AIs made confident identifications of objects that were not only incorrect, but that looked vastly different from the purported object to human eyes. The AIs interpreted images that to the human eye looked like static or abstract wavy lines as animals or other objects, and asserted greater than 99.6% confidence in their estimation.[18]

[17] Andy Greenberg, "Hackers Remotely Kill a Jeep on the Highway – With Me in It," *Wired*, July 21, 2015, http://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/.

[18] Anh Nguyen, Jason Yosinski, Jeff Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *Computer Vision and Pattern Recognition*, IEEE (2015), http://arxiv.org/pdf/1412.1897v4.pdf.

IMAGE IDENTIFICATION BY LEARNING NEURAL NETWORK (>99.6% CONFIDENCE)[19]



[19] ibid.

The problem is not that the neural network misidentified some images. Some mistakes are inevitable and the same neural network generally performs well. The problem is that the way in which the network misclassified these images is completely alien to humans. This significantly complicates a human's ability to predict how the neural network might classify objects, since its internal cognitive processes are vastly different from a human's.[20]

Neural networks and complex rule-based autonomous systems share a common problem: The complexity of the system can make its operation opaque to trained users or even its designers, and as a result the system sometimes can behave unexpectedly.

Predicting the boundaries of behavior

A common misunderstanding is that complex autonomous systems are "unpredictable." That is not necessarily the case. The behavior of the autonomous system may be predictable and safe under most operating conditions. The problem is an inability to confidently verify the behavior of the system under all possible operating conditions. As a result, as systems increase in complexity, human operators may have greater uncertainty regarding the conditions under which the system will fail. This makes it more difficult for the human operators to avoid failures. When combined with autonomous systems that have a high damage potential if they fail, the result could be significant risk.

[20] David Berreby, "Artificial Intelligence is Already Weirdly Inhuman: What Kind of World is Our Code Creating," Nautilus, August 6, 2015. http://nautil.us/issue/27/dark-matter/artificial-intelligence-is-already-weirdly-inhuman.

V. Autonomous Weapons and Unintended Engagements

Autonomous weapons are a special kind of autonomous system. In autonomous weapon systems, the task being performed is selecting and engaging targets on the battlefield. Once activated, an autonomous weapon will select and engage targets on its own. It selects targets according to pre-programmed criteria written by humans, but human operators have not chosen the specific targets to be engaged.

The risk in using an autonomous weapon is that it selects and engages targets other than what the human operator intended. This could result in fratricide, civilian casualties, or unintended escalation in a crisis. The U.S. Department of Defense policy on autonomy in weapons characterizes this undesirable outcome as an "unintended engagement," which it defines as: [21]

> **Unintended engagement:** The use of force resulting in damage to persons or objects that human operators did not intend to be the targets of U.S. military operations, including unacceptable levels of collateral damage beyond those consistent with the law of war, ROE, and commander's intent.[22]

A number of key variables determine the risk—the probability and consequences—of an unintended engagement:

- **What is the inherent hazard of the system?** What is the level of armament of the weapon system and against what kinds of targets? An autonomous weapon employing 2,000-pound bombs has a greater inherent hazard than one armed with a taser. The type of target against which the weapon operates is also a consideration. All things being equal, an anti-personnel autonomous weapon that targets people poses a greater inherent hazard to civilians than an anti-vehicle or anti-materiel autonomous weapon, such as one that only targets tanks or radars.

- **What is the time delay between when any failures occur and potential corrective action?** Semi-autonomous weapon systems allow for real-time feedback to human operators before they approve additional engagements, so that human operators can halt the system if it is not performing appropriately. Supervised autonomous weapons also give human operators the ability to intervene and halt the weapon, but the time from failure to corrective action depends on any time delays in the system and the timeliness of the human operator's response. Fully autonomous weapons, by definition, would be those that have no ability to undertake corrective action while it is performing the task (presumably because of a lack of

[21] The purpose of the policy directive 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." U.S. Department of Defense, *Department of Defense Directive 3000.09, Autonomy in Weapon Systems* (2012), 1, http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

[22] ibid, 1.

communications), although the weapon could be corrected before being used on additional missions.

- **What is the damage potential of the system?** The damage potential is the amount of damage that an autonomous weapon performing inappropriately could do before corrective action could be taken. This depends on: its inherent hazard; the time between failure and corrective action; the speed of engagements; the weapon's magazine depth (total amount of ordnance it is carrying); and its geographic reach. An intercontinental bomber carrying 40,000 pounds of ordnance has a much higher damage potential than a pistol-armed quadcopter, even if both are fully autonomous weapons.

- **What is the aggregate damage potential of *all* autonomous weapon systems of that type in operation at one time?** Because a software flaw in any one system is likely to be replicated across all other identical systems, if one autonomous weapon is susceptible to hacking or other failures, then others are likely to be as well. Militaries must consider the aggregate damage potential of all autonomous weapons of that type in operation at one time if they begin to fail.

Semi-autonomous vs. autonomous weapons

To examine the operational risks associated with autonomous weapons, consider four types of possible offensive anti-radar weapons, varying by degree of autonomy and type of vehicle (single-use munition vs. recoverable platform).

POSSIBLE OFFENSIVE ANTI-RADAR WEAPONS

	Semi-Autonomous Weapon System	**Fully Autonomous Weapon System**
Munition (single-use)	Anti-radiation homing missile (e.g., HARM)	Anti-radiation wide area loitering munition (e.g., Harpy)
Platform (recoverable)	Semi-autonomous radar hunting drone (hypothetical)	Fully autonomous radar hunting drone (hypothetical)

Even if these systems were to use the same type of sensor and targeting algorithm, a failure that led them to erroneously engage friendly forces or civilian objects could have very different consequences.

- **Semi-autonomous munition (homing missile):** The high-speed anti-radiation missile (HARM) is a fire-and-forget weapon that, once launched, homes in on enemy radars and destroys them. It is autonomous in the sense that, once launched, no further interaction is required by the human operator. However, its field of view and ability to loiter are

constrained such that, provided the operator has employed it properly, it will engage only the specific radar that the human operator has intended. The missile is limited in its ability to search for targets. In order to be used effectively, the missile has to be launched at a known or likely target.

- **Fully autonomous munition (wide area loitering munition):** The Harpy is a loitering anti-radar weapon that searches for radars over a wide area and, once it finds them, kamikazes into them. While the HARM has a very limited time of flight—approximately four and a half minutes—the Harpy, by contrast, can stay aloft for over two and a half hours.[23] This gives it the ability to search for enemy radars over a wide area. Operators employing the Harpy do not need to know in advance specifically where enemy radars will be, only the general location of likely or suspected radar sites. Similar to the HARM, however, the Harpy is a single-use munition. It can engage only one radar and cannot return.

- **Semi-autonomous platform (semi-autonomous drone):** A number of nations are developing next-generation stealth combat drones. The same anti-radiation seekers used on the HARM or Harpy could, in principle, be used to design a semi-autonomous radar-hunting drone. Like a Harpy, it would loiter over a wide area searching for enemy radars. However, unlike the Harpy, once it found an enemy radar, it would radio target data back to remote human controllers for approval. This could include an image of the target for confirmation as well as its surrounding area for proportionality considerations. This could be done at relatively low bandwidth.[24] Once authorized by the human controller, the drone would launch a HARM-like weapon at the specific enemy radar that was authorized. The entire engagement process would essentially be the same as it is today with human-inhabited ("manned") aircraft, where automation is often used to assist in target identification, except in this case the human pilot would be remote.

- **Fully autonomous platform (fully autonomous drone):** Alternatively, militaries could decide to take the human out of the loop entirely, building a fully autonomous radar-hunting drone.[25] This might be desired if they believed they would not have any communications in a contested environment, even short-range communications to human controllers in nearby aircraft. Similar to the Harpy, the fully autonomous drone would have the ability to search over a wide area for likely and suspected enemy radars and strike them without further approval. Unlike the Harpy, however, it would have the ability to strike multiple radars in a single mission and then return to base.

[23] Robert O'Gorman and Chris Abbott, "Remote Control War: Unmanned combat air vehicles in China, India, Israel, Iran, Russia, and Turkey" (Open Briefing, September 2013), 75; and "AGM-88 HARM Missile," United States Navy, February 20, 2009, http://www.navy.mil/navydata/fact_display.asp?cid=2200&tid=300&ct=2.

[24] Paul Scharre, "Yes, Unmanned Combat Aircraft are in the Future," War on the Rocks, August 11, 2015, http://warontherocks.com/2015/08/yes-unmanned-combat-aircraft-are-the-future/.

[25] No nation has stated that they intend to build fully autonomous weapons.

Even though all of these weapons engage enemy radars and could rely on the same targeting seeker and algorithm, the risk in employing these weapons is quite different. Consider the consequences with each system if the targeting algorithm failed such that it misidentified friendly or civilian objects as enemy targets.

- In the case of the **semi-autonomous HARM**, using the weapon is not risk-free, but the risk is fairly limited. In the event of a failure where the seeker misidentified friendly or civilian objects as enemy targets, those objects would have to be within the seeker's acquisition basket for the short period of time in which the seeker is active. Because the system's autonomy is constrained in time and space, its ability to search for erroneous targets is limited. Furthermore, because human authorization is required for each HARM launch, in the event of an unintended engagement, the human operator can easily halt subsequent engagements by not firing additional HARMs.

- A **fully autonomous Harpy**, on the other hand, has much greater freedom of action in space and time. The same failure mode on a Harpy—misidentifying friendly or civilian objects as enemy targets—could have greater consequences than on a HARM. Because the Harpy has a wider area and a longer period of time in which to search for targets, the odds that it finds an erroneous target are higher. Thus, the damage potential of a Harpy is greater because it has a greater freedom of action in space and time, even if the size of the warhead (inherent hazard) is the same. Because the Harpy is a single-use munition and each launch is a discrete decision, the human operator does have the ability to halt subsequent engagements by not firing additional Harpies. However, it may not be possible to recall Harpies already in the air, which may be a problem given their long loiter time (more than 2.5 hours) and wide geographic coverage.

- A **semi-autonomous radar-hunting drone** would be constrained in its freedom of action, similar to a HARM, even though it can loiter over a wide area. Human controllers would approve strikes against specific enemy radars, but since strike authority would be constrained in time and space against a specific target, the damage potential in the event of a targeting failure would be more limited. If the system began to erroneously target friendly radars or civilian objects, the human controller would still have to approve each target before engagement. Even if one missile went awry, it would be only a single missile with a limited seeker field of view and loiter time, similar to a HARM. The human operator could then correct for the system's limitations in subsequent engagements, or simply halt all engagements. The damage potential is therefore far more limited, like a HARM.

- A **fully autonomous radar-hunting drone** would have greater damage potential than a semi-autonomous drone. If it began to fail, it could continue engaging inappropriate targets until it exhausted its ammunition. It also would have greater damage potential than a Harpy, since it could carry multiple missiles. While a malfunctioning Harpy could only engage, at most, one wrong target, a malfunctioning drone could engage many incorrect targets before

> running out of ammunition. In addition, depending on the drone's design, its geographic reach and endurance could be even larger than that of a loitering munition, staying aloft for 8 to 24 hours and potentially roaming thousands of miles.

The decision the human operator makes in launching these weapons is also different in each case. Because the semi-autonomous HARM must be launched at a specific known or likely enemy radar in order to be effective, the human operator must make a determination on the military necessity and likely collateral damage for that specific target. This means the human operator can take into account other factors that the missile itself may not be able to account for, such as whether the radar is within an urban area and its proximity to civilians.

For a fully autonomous loitering munition like the Harpy or a radar-hunting drone, this is not the case. The human operator is only making a determination about the military necessity and likely collateral damage associated with targeting radars in this wide area *in general*.[26] The increased freedom of action of the autonomous weapon reduces the human operator's ability to account for specific circumstances where the weapon's use *in that particular instance* may not be appropriate.[27]

Finally, because a drone is recoverable, it is likely to be used differently by human operators than a loitering munition like the Harpy. Since the Harpy cannot be recovered, human operators cannot send it on patrol. Launching a Harpy is equivalent to a decision to use lethal force, even if the human operators cannot know exactly which specific enemy radar will be destroyed. A drone, on the other hand, could be used to patrol an area prior to the initiation of hostilities. Doing so with a fully autonomous drone would be tantamount to not only delegating target selection to an autonomous weapon, *but potentially the decision to initiate hostilities itself*.

While countries may decide not to send armed and activated fully autonomous weapons on patrol in peacetime, it is entirely plausible that they might be used to patrol a conflict area during a period of heightened tension. In fact, they might be preferred for a variety of reasons. Without a human on board, the platform could be sent into more dangerous areas without risking a human life. Autonomous target selection and engagement could be desired for quicker reactions to enemy attacks or to give the drone the ability to defend itself if its communications were jammed. The result, however, could be unintended escalation if the system engaged an otherwise legitimate enemy target but in a situation where the human operator did not intend an engagement. In a crisis, the consequences could be severe.

---

[26] Even though the weapon system is autonomous and selects its own targets, the human employing the weapon is still responsible for ensuring the lawfulness of the engagement before using the weapon.

[27] This distinction between a semi-autonomous weapon and an autonomous weapon is also particularly important from a legal and ethical perspective, since the commander employing the autonomous weapon it still responsible for the consequences arising from its operation. However, these issues are beyond the scope of this paper.

Assessing the risk of autonomous weapons

Fratricide, accidents, and civilian casualties are unfortunate realities of war. Humans make mistakes, and no weapon system can be expected to operate 100 percent perfectly. Some weapons are more dangerous than others, however, necessitating greater safeties and higher levels of approval for their use.

Because of their higher damage potential in the event of a failure, autonomous weapons have a qualitatively different level of risk than equivalent semi-autonomous weapons. Without a human in the loop to take corrective action, the consequences of a failure with an autonomous weapon could be much more severe. A failure with an autonomous weapon could lead to multiple unintended engagements across a wide geographic area until the platform exhausts its ammunition.[28] The actual damage potential of any given autonomous weapon depends on the size of the warhead being used, the platform's magazine capacity, and its range and loiter time. But the key insight is that a higher damage potential related to an equivalent semi-autonomous weapon *is intrinsic to the nature of an autonomous weapon system*.

As militaries weigh these risks, they must consider the aggregate damage potential resulting from potentially multiple systems falling victim to the same failure mode at the same time. Even if the likelihood of a failure in an autonomous weapon is low—potentially lower than a human-controlled weapon—the consequences of a fleet of autonomous weapons failing in a manner that led to multiple unintended engagements could be catastrophic: significant civilian casualties, unintended escalation in a crisis, or mass fratricide across the battlefront.

This is a key difference between autonomous and semi-autonomous weapons: While humans sometimes make mistakes that lead to fratricide and civilian casualties, humans tend to be idiosyncratic. Different people in the same situation might come to very different conclusions. One human mistake can lead to a large number of casualties, but it is extremely unlikely that thousands of humans across the battlefield will make precisely the same mistake in precisely the same way. Human diversity and heterogeneity poses an inherent resilience against mass failures.

The increased damage potential of autonomous weapons relative to semi-autonomous ones significantly complicates thinking about the risk associated with employing them. Autonomous weapons may have operational advantages in some scenarios. They could enable faster reactions to enemy attacks. In communications-denied environments, they could allow uninhabited platforms to defend themselves or attack emerging targets. For this reason, militaries will want to carefully consider ways to minimize the likelihood and consequences of failures that could lead to unintended engagements. Better test and evaluation, training, software verification and validation, and

[28] An autonomous weapon that is failing in this manner could even cause significant problems without any actual inappropriate engagements. Simply the existence of an autonomous weapon that is performing inappropriately and not responding to commands could significantly complicate military operations, akin to a "mad guard dog" that no longer responds to commands. This would be analogous to the problem of mining, but could potentially lead to a condition of "accidental mining" via autonomous weapons and over a much wider area. Thanks to John Borrie for making this point.

cybersecurity can all help to minimize the likelihood of failures.[29] However, these efforts are complicated by two unfortunate realities: the inevitability of failures in complex systems; and adversarial risk from hacking, spoofing, or behavioral manipulation of autonomous systems.

[29] For example, the U.S. Department of Defense policy on autonomy in weapons contains an exhaustive set of requirements designed to reduce the probability and consequences of unintended engagements. U.S. Department of Defense, *Department of Defense Directive 3000.09, Autonomy in Weapon Systems*, Appendices A-B.

VI. The Inevitability of Failure: Complex Systems and Normal Accidents

It is tempting to think that risk can be designed out of complex systems, but this is not the case. Risk can be reduced but never entirely eliminated. It is impossible to account for all of the possible interactions that can occur in sufficiently complex systems.

As a result, complex systems are potentially vulnerable to system failures due to components interacting in unexpected or nonlinear ways. These can stem from interactions within the system itself, with human operators, or with its environment.

When there is sufficient "slack" in the system in terms of time between interactions and the ability for humans to exercise judgment, bend or break rules, or alter the system's behavior, then these accidents can be managed without catastrophe. However, when components of a system are "tightly coupled" such that failures can rapidly cascade from one subsystem to the next with little slack in between to absorb and react to failures, accidents can become inevitable or even "normal."

"Normal accident" theory suggests that in tightly coupled complex systems, such as modern military weapon systems, accidents may not be likely but are inevitable over a long enough time horizon. Moreover, because complex systems often interact in ways that are unanticipated and nonlinear, the rate of such accidents often cannot be predicted accurately in advance. System designers may not be able to accurately assess the probability of a failure, and hidden failure modes may lurk undetected.[30]

These risks can occur in complex systems as wide ranging as jet airliners, spacecraft, or nuclear power plants. While they can be mitigated through various safety measures, they cannot be eliminated. Complex systems like airplanes, spacecraft, or nuclear power plants can be made safer, but never 100 percent safe.

Three Mile Island

The textbook example of a normal accident is the Three Mile Island nuclear power plant partial meltdown in 1979. The Three Mile Island incident highlights many of the features of normal accidents that are applicable in a broad range of complex systems, including autonomous weapons. Chief among these are:

- A multitude of possible causes of failures, including human error, equipment malfunction, and poor system design
- Surprising and unanticipated interactions between various elements of the system

[30] Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (Princeton, NJ: Princeton University Press, 1999). See also Borrie, "Safety aspects of 'meaningful human control': Catastrophic accidents in complex systems."

- The sheer complexity of the system itself acting as an obstacle to human operators' understanding of its behavior, making it sometimes incomprehensible
- Tight coupling between various elements of the system causing simple failures to rapidly cascade to catastrophe with little time or flexibility for human operators to intervene and react to circumstances

The trouble began in the early hours of March 28, 1979, with a seemingly inconsequential failure in a non-critical piece of equipment. [31]The initial problem was a leaky seal in the water cooling system (equipment malfunction). Moisture from the leaky seal got into an unrelated system, causing it to shut off two water pumps vital for cooling the nuclear reactor (unanticipated interaction between unrelated components). Once this occurred, automated safeties kicked in, shutting down the turbines that generate electric power and turning on emergency water pumps to cool the reactor. However, a valve needed to allow water to flow through the emergency cooling system had been left closed (human error). Human operators managing the reactor were unaware of this problem because an indicator light on their control panel showing the valve's position was obscured by a repair tag for another, unrelated system (poor system design, unanticipated interaction between unrelated components).

Without water circulating through the emergency cooling system, the reactor began to overheat. The excess heat activated another automatic safety and the reactor "scrammed," dropping graphite control rods into the reactor core to absorb neutrons and stop the reaction. This should have been sufficient to halt the accident. At this point, the nuclear chain reaction had stopped. However, the core was still producing heat and needed continuous water flow to keep it cool.

Both the primary and emergency cooling systems had already failed, however. Without cold water flowing through the cooling system, the water trapped in the core began to heat up and the pressure began to rise. Excessively high pressures in the reactor core are dangerous because they can crack the containment vessel, releasing radiation. In response to the rising pressure, another automatic safety kicked in: an automated pressure-release valve. It was designed to automatically open to let off steam if the pressure got too high, then close once the pressure had returned to normal levels.

The automated pressure-release valve opened as designed, but did not close (equipment malfunction). The valve's indicator light also failed (equipment malfunction), so operators were not aware that the valve failed to close. With the valve stuck open, too much steam was released, and a third of the water in the core escaped. Water is crucial to cooling the still-hot nuclear core, so another automatic safety kicked in: an emergency water cooling system. The plant's operators also activated another emergency water cooling system.

[31] The accident description is taken from Perrow, *Normal Accidents*, 15-31. See also, United States Nuclear Regulatory Commission, "Backgrounder on the Three Mile Island Accident," http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html#summary.

All of this occurred within 13 seconds after the start of the accident. Multiple automatic safeties kicked in and the plant's human operators reacted swiftly. However, none of these actions addressed the root cause of the problem: 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. The indicators that would have told the plant's operators about these problems similarly failed, one obscured by a repair tag and another simply malfunctioning. In describing the accident, Charles Perrow points out, "*The operators could have been aware of none of these.*" [emphasis original][32]

This highlights an important point about complex systems: Their sheer complexity can make their behavior seem incomprehensible to human operators. The human operators at Three Mile Island were attempting to understand the system through a dizzying array of gauges and indicators—some 2,000 alarms in the control room.[33] The system was so complicated that human operators were guessing at the causes of the accident. They could see temperature and pressure gauges spiking, but did not understand what was causing these problems inside the reactor. Moreover, some of their control indicators were malfunctioning, but the plant's operators could not possibly know which ones to trust. As a result, the operators did not discover that the emergency water cooling valve was improperly closed until eight minutes into the accident and did not discover that the pressure-release valve was stuck open until two hours and 20 minutes into the accident. 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 incomprehensibility of complex systems is a key feature of normal accidents and a major obstacle in human operators' ability to intervene and regain control of a complex highly automated or autonomous system. Even when human operators are operating in a supervisory control mode where they are, in principle, able to intervene, they can't possibly take the right corrective actions if they don't understand how the system is behaving and why. This can mean that the time from initial failure to the right corrective actions being applied may be longer than it otherwise might have been. In the interim, human operators may, through no fault of their own, take actions that actually exacerbate the problem. In the case of Three Mile Island, engineers were able to manage the disaster without any loss of life, although cleanup cost $1 billion.[34] In the case of Air France Flight 447, by the time the pilots discovered their error, it was too late and the plane crashed, killing everyone onboard.

Each individual failure that led to the catastrophe at Three Mile Island—equipment malfunctions, human errors, and poor design choices—was simple enough and could, in principle, have been prevented. The problem—and the essence of normal accidents—was that these seemingly minor failures combined in unpredictable and nonlinear ways to produce dramatic consequences. Chief among these is the interaction among unrelated components. In reference to this element of Three

[32] Perrow, *Normal Accidents,* 22.
[33] William Kennedy, interview, 2015.
[34] "14-Year Cleanup at Three Mile Island Concludes," *New York Times,* August 15, 1993, http://www.nytimes.com/1993/08/15/us/14-year-cleanup-at-three-mile-island-concludes.html.

Mile Island, Perrow states: "Here we have the essence of the normal accident: the interaction of multiple failures that are not in a direct operational sequence." [35] The unexpected interaction of these components makes predicting these failures in advance essentially impossible. In complex systems, it is simply not feasible to map all of the possible interactions of the system with itself, its operators, and its environment. (This problem is even further exacerbated in competitive situations where adversaries will try to seek out weaknesses in the system to hack or exploit.)

Unexpected interactions may be manageable in loosely coupled systems where the human operators have time and flexibility to respond to unforeseen events. One of the major advantages of humans over automation is the ability of humans to adapt to unanticipated problems and arrive at novel solutions. The ability of human operators to adapt is severely curtailed, however, when a system is tightly coupled so that one failure can rapidly cascade to the next. When this is combined with a system whose complexity makes it largely incomprehensible, human operators who are nominally supposed to be in control may find themselves instead merely along for the ride, "supervising" a complex system that has spun out of control.

Space accidents: *Apollo 13*, *Challenger*, and *Columbia*

These features of normal accidents in complex systems—unexpected interactions, tight coupling, and the incomprehensibility of the system itself—come up time and again in complex systems in a variety of high-risk settings. 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, and this was 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.[36] Nevertheless, the system was incomprehensible enough that vital time was lost as the astronauts and ground control experts pored over the various instrument readings and rapidly-cascading electrical failures before they discovered the root cause.[37] Similar to Air France Flight 447 and Three Mile Island, the complexity of the system limited the human operator's ability to understand what was happening and regain effective control, extending the time from failure to when the operators could undertake corrective action.

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

[35] Perrow, *Normal Accidents,* 23.
[36] For a very brief summary of the incident, see National Aeronautics and Space Administration, "Apollo 13," https://www.nasa.gov/mission_pages/apollo/missions/apollo13.html. 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," June 15, 1970, http://nssdc.gsfc.nasa.gov/planetary/lunar/apollo_13_review_board.txt.
[37] Perrow, *Normal Accidents,* 271-278.

crews.[38] While these accidents had discrete causes that could be addressed in later designs (faulty O-rings and falling foam insulation, respectively), it is the inability to anticipate these specific failures in advance that makes continued accidents inevitable, if rare. Earlier this year, for example, the private company Space-X had a rocket blow up on the launch pad due to a strut failure that had not been previously identified as a risk.[39]

The 2011 Fukushima Daiichi meltdown

Nuclear power similarly has greatly improved in safety since Three Mile Island, and indeed in large part because of it, as well as the 1986 Chernobyl disaster. The 2011 meltdown of the Fukushima Daiichi nuclear power plant in Japan points to the challenge in safely operating complex systems with high damage potential, however. Unlike Three Mile Island, the Fukushima meltdown, which was far more severe, was caused not by internal failures but by a massive external shock to the system from its environment: a 9.0 magnitude earthquake off the coast of Japan, the largest recorded earthquake ever to hit Japan.

The Fukushima Daiichi plant was hardened against earthquakes and against a loss of power. The earthquake's shock waves did not damage the plant, and when the plant lost power, emergency procedures worked correctly. The three reactors that were active at the time automatically scrammed, inserting control rods to stop the nuclear reaction. Backup diesel generators came online to restore power to vital water cooling systems.

However, the backup diesel generators were in low-lying areas. They were protected from flooding by 30-foot-high flood walls, but the earthquake-induced tsunami waves topped 40 feet. The waves crested the walls and swamped the backup generators, taking 12 of Fukushima Daiichi's 13 backup generators offline. Coupled with the loss of the power from the electrical grid, the result was that the plant lost the ability to pump water to cool its reactors. 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.[40]

The plant had been designed to fail safely in the event of an earthquake or loss of power by reverting to backup systems. Additionally, it had 30-foot-high flood walls to hold back massive waves. What the plant was not prepared for was an event that both took out the primary grid power *and* triggered flooding that topped the flood walls and swamped the backup electrical generators. This is referred to as a "common-mode failure," a situation where a single event causes multiple seemingly-independent systems to fail. The Fukushima Daiichi plant could handle flooding, earthquakes, or

[38] On *Challenger*, see National Aeronautics and Space Administration, "Report of the Presidential Commission on the Space Shuttle Challenger Accident," June 6, 1986, http://history.nasa.gov/rogersrep/51lcover.htm. On the *Columbia* accident, see National Aeronautics and Space Administration, "Columbia Accident Investigation Board, Volume 1," August 2003, http://spaceflight.nasa.gov/shuttle/archives/sts-107/investigation/CAIB_medres_full.pdf.

[39] Space-X, "CRS-7 Investigation Update," July 20, 2015, http://www.spacex.com/news/2015/07/20/crs-7-investigation-update.

[40] Phillip Y. Lipscy, Kenji E. Kushida, and Trevor Incerti, "The Fukushima Disaster and Japan's Nuclear Plant Vulnerability in Comparative Perspective," *Environmental Science and Technology* 47 (2013), http://web.stanford.edu/~plipscy/LipscyKushidaIncertiEST2013.pdf.

loss of power. It was not prepared for a massive earthquake off the coast that triggered all of these simultaneously. This highlights an additional possible cause of normal accidents: unexpected interactions with the environment.

Accidents are inevitable, even "normal," in tightly-coupled complex systems

It is easy to say now that that Fukushima Daiichi should have had higher flood walls or elevated platforms for the backup generators. In retrospect, all of these disasters—Three Mile Island, *Apollo 13*, *Challenger*, *Columbia*, the Space-X rocket explosion, Air France Flight 447, and Fukushima Daiichi—could have been prevented if those specific failure modes had been anticipated. But they weren't. This was not because the organizations operating these systems were sloppy or lazy. Safety is the watchword in aviation, space travel, and nuclear power plant operations. The organizations responsible for designing and operating these systems are composed of highly trained engineers from advanced industrialized nations, and safety is a major focus of their efforts. And many, many failures are anticipated and prevented. Accidents with nuclear power, commercial air travel, or even space travel are rare. But they can never be prevented entirely. Over a long enough time horizon, unanticipated system interactions are inevitable.

Normal accidents in military systems: the Patriot fratricides

Many modern military systems exhibit the same degree of complexity as commercial airplanes, spacecraft, or nuclear power plants. When these systems are tightly coupled such that one failure can directly cascade to other components, the systems are similarly at risk of normal accidents.

During the 2003 invasion of Iraq, the U.S. Patriot air defense system shot down two friendly aircraft. The causes behind the Patriot fratricides illustrate how normal accidents also can occur in military systems.

The first fratricide occurred on March 24, 2003, when a U.S. Patriot battery shot down a British Tornado aircraft, killing the crew. The Patriot's automation misidentified the Tornado as an anti-radiation missile (which would have been a valid target for engagement). An identification friend or foe (IFF) system, which allows friendly military aircraft to identify themselves to friendly radars, should have prevented this incident. However, it "performed very poorly," according to a Defense Science Board Task Force report on the incident. The Task Force further remarked that the poor IFF performance was a known problem:

> 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.[41]

[41] 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*, 20301-3140 (January 2005), http://www.acq.osd.mil/dsb/reports/ADA435837.pdf.

These two failures (target misidentification and IFF failure) were not enough to result in fratricide. The Patriot was operating in a semi-autonomous mode and required human approval for each engagement. However, the human operator accepted the Patriot's (incorrect) identification of the aircraft as an anti-radiation missile and authorized the engagement. Afterward, a lengthy internal Army investigation criticized the Patriot community culture for "trusting the system without question."[42] According to Army researchers, 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."[43]

In the first Patriot fratricide, three separate, independent systems failed: automated target identification, IFF, and the human in the loop. A failure in any one or even two of these systems would have been manageable, but the failure of all three at the same time was not. The target misidentification may not have been preventable, but the IFF failure and operator automation bias were. Inadequate system design and operator training were at fault.

The second Patriot fratricide occurred a little over a week later, on April 2, when a Patriot shot down a U.S. Navy F/A-18C Hornet fighter, killing the pilot. In this incident, the cause was more far more complex.

Following the first fratricide, Patriot systems were kept in a standby mode to prevent automated engagements. The second fratricide began when the Patriot identified an incoming track from a ballistic missile. Later, it was determined that the track was false, likely the result of electromagnetic interference, possibly due to employing radars in a non-standard configuration in the field.[44]

However, the operators were not aware that the missile track was false. In response to the reported incoming ballistic missile, the human operator brought the system to a "ready" status to prepare for an engagement. However, the Patriot battery was in an auto-fire, not a semi-autonomous, mode. This meant that once it came to ready, it was authorized to engage any active threats. (The operator either was unaware that the system was in an auto-fire mode or had forgotten in the heat of the moment that bringing the system to ready in this situation was tantamount to an order to fire.) Once the system came to ready, the Patriot battery fired. For reasons that still remain unclear to Army investigators, the missile shot down an F-18 in the vicinity. (It is possible that the F-18 was simply in the wrong place at the wrong time and when the missile deployed its onboard seeker, it tracked onto the nearby F-18.)[45]

---

[42] John K. Hawley, "Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons," *Army Research Laboratory*, September 2007, http://www.arl.army.mil/arlreports/2007/ARL-SR-0158.pdf.

[43] John K. Hawley, "Not by Widgets Alone: The Human Challenge of Technology-intensive Military Systems," *Armed Forces Journal*, February 1, 2011, http://www.armedforcesjournal.com/not-by-widgets-alone/. Patriot operators now train on this and other similar scenarios to avoid this problem of unwarranted trust in the automation.

[44] David Talbot, "Preventing 'Fratricide,'" *MIT Technology Review*, June 1, 2005, http://www.technologyreview.com/article/404191/preventing-fratricide/page/3/.

[45] John Hawley, personal correspondence.

Again, multiple circumstances contributed to the fratricide. The initial failure was caused by an unanticipated interaction with the environment and/or nearby military radars (electromagnetic interference resulting in the false missile track). Inadequate system design and/or operator training created a situation where the operators did not understand how their actions would affect the system (that bringing the system to ready while in an auto-fire mode would cause it to fire). These combined to launch a missile against a false target. Why the missile tracked onto the nearby F-18 is still unknown.

The Patriot fratricides demonstrate many of the same features of normal accidents in other non-military complex systems. A confluence of failures—some human and some machine, some anticipated in advance and some novel—contributed to the fratricides. The complexity of the system contributed to human operators' misperceiving or misunderstanding the system's behavior, in some cases taking inappropriate actions. The Patriot fratricides also demonstrate the dual-edged sword of increased automation in complex systems. While automation can be useful in handling many routine problems, it can lead to a case of automation bias, where operators trust the automation too much. (Conversely, trusting the automation too little and intervening when it isn't necessary can also lead to disaster, as in the case of the Air France 447 crash.)

Like many normal accidents, the Patriot fratricides could be considered freak occurrences. But this is, of course, the essence of normal accidents. In complex tightly coupled systems given enough operational use, unanticipated system failures are bound to occur. They may be rare, but they are inevitable, even "normal." These two fratricide incidents, when placed in the broader context of the Patriot's operational history during Operation Iraqi Freedom, demonstrates this problem even more sharply.

The two fratricides comprised 18 percent of the Patriot's 11 total engagements during Operation Iraqi Freedom, an "unacceptable" fratricide rate, according to Army investigators.[46] When considered as a fraction of total possible fratricides, the Patriot fratricide rate was quite low. Sixty Patriot batteries were deployed during the Iraq invasion, during which coalition aircraft participated in 41,000 sorties. As the Defense Science Board Task Force pointed out, with this many units in operation, "the possible Patriot-friendly aircraft observations were in the millions and even very-low-probability failures could result in regrettable fratricide incidents."[47] Thus, even a very low probability of failure can result in an unacceptably high number of fratricides if the number of possible interactions with friendly systems is high, as it was in the case of the 2003 Iraq invasion.

This paradox—that even very low probability events can become effectively inevitable given enough exposure—is what makes unlikely accidents "normal" in complex systems.[48] As Perrow explains:

---

[46] The problems that led to the fratricides have since been corrected in the Patriot systems and operator training.

[47] 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.*

[48] Perhaps the best illustration of this feature of low-probability events is the lottery. The odds of any individual winning the lottery are astronomically low. But someone wins because enough people buy tickets.

> [E]ven 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. These systems are currently too complex and tightly coupled to prevent accidents that have catastrophic potentials. We must live and die with their risks, shut them down, or radically redesign them.[49]

Choosing to operate tightly coupled complex systems in high-risk environments thus entails accepting the risk that, over a long enough time horizon, failures are bound to occur.

[49] Perrow, *Normal Accident*s, 354.

VII. Adversarial Risk: Normal Accidents in Competitive Environments

While military systems can exhibit the same kind of complexity that leads to normal accidents, they differ from nuclear power plants, airliners, or spacecraft in one crucial way: Military systems operate in a competitive environment against an adversary. Everyone involved in the operation of a spacecraft or nuclear power plant is trying to get the system to operate safely. There is no enemy out to sabotage its operation. However, for militaries, adversaries are not merely incidental to the system's operation, they are its very reason for being.

This added competitive dimension increases the possible ways in which failures can occur. These include:

- Incomplete information
- An accelerated pace of interactions
- Unanticipated interactions between adversarial systems
- Hacking
- Spoofing (sending false data)
- Behavioral hacking (exploiting predictable behaviors)

"The Man Who Saved the World" [50]

Perhaps the most frightening near-accident of all time comes from a Cold War–era automated nuclear warning system, in which a human in the loop may have saved humanity from destruction.

On September 26, 1983, a Soviet automated missile alert system reported the launch of five U.S. intercontinental ballistic missiles at the Soviet Union. Per Soviet doctrine, the military officer on duty was required to report the attack to higher headquarters. The officer on duty, however, Lieutenant Colonel Stanislav Petrov, judged that a U.S. first strike consisting of only five missiles was nonsensical and likely an error in the new computer system. Rather than report a U.S. attack, he reported a system malfunction. Later, he was found to be correct—Soviet satellites were picking up false positives of "missile launches" from sunlight reflecting off of clouds (unanticipated interaction with the environment).

It is unclear how the information would have been treated at Soviet higher headquarters if Petrov had reported the incoming missiles as a genuine attack. The incident came at a period of high Cold War tensions and just before a major NATO military exercise, Able Archer 83, which some Soviet leaders feared was masking preparation for a surprise NATO attack. A former CIA analyst has

[50] This incident is depicted in the 2014 documentary, *The Man Who Saved the World*, directed by Peter Anthony, produced by Jakob Staberg, 2014, http://www.imdb.com/title/tt2277106/.

described the incident as "probably the most single dangerous incident of the early 1980s."[51] The counterfactual of what would have occurred if Petrov had reported the incident up the chain as an attack can only be surmised. But it is clear that the inclusion of a person in the loop for the decision on how to respond to the missile warnings permitted human judgment to exercise a more nuanced understanding of the facts, including the broader context, which in this case helped defuse a potentially existential danger.[52]

The 1983 Stanislav Petrov incident was caused by an unanticipated interaction with the environment, but the risk was exacerbated by a number of conditions unique to adversarial environments—namely, a lack of complete information and a competitive dynamic that shortened decision timelines. The Soviets very well could have called U.S. military leaders at the Pentagon via a hotline that was established following the Cuban Missile Crisis to ask if they had fired their missiles. However, whether they would have trusted U.S. officials is another matter entirely. If the United States had launched its missiles, obviously they would not tell the Soviets, making their answer effectively irrelevant. All of these considerations were compounded by the fact that the Soviets had only minutes to decide whether to launch a counterattack or not, if the incoming missiles were real.

Flash crash

More recently, automated stock trading algorithms offer an example of the risks of autonomous systems interacting in complex, competitive environments and at speeds exceeding human reaction times. On May 6, 2010, the Dow Jones Industrial Average experienced a "flash crash" where it lost nearly 10 percent of its value in a matter of minutes.

A U.S. Securities and Exchange Commission (SEC) report following the incident determined that the crash was initiated by an automated stock trade (a "sell algorithm") executing a large sale unusually quickly. This caused a sale that normally would have occurred over several hours to be executed within 20 minutes.[53] This sell algorithm then interacted with high-frequency trading algorithms to cause a rapid price drop (unanticipated interaction with competitive autonomous systems).[54]

---

[51] David Hoffman, "I Had a Funny Feeling in My Gut," *The Washington Post*, February 10, 1999, http://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/shatter021099b.htm.

[52] It would be reassuring to think of the Stanislav Petrov incident as an isolated occurrence. However, have been an alarming number of near-nuclear incidents over the past 60 years. See Patricia Lewis et al., "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy," *Chatham House Report,* April 2014, https://www.chathamhouse.org/sites/files/chathamhouse/field/field_document/20140428TooCloseforComfortNuclearUseLewisWilliamsPelopidasAghlani.pdf. See also Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1993).

[53] The Sell Algorithm could have been programmed take into account price and time in executing the sale, but wasn't. Whether this was due to poor training, opaque system design, or simple human error is not clear. U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission, *Findings Regarding the Market Events of May 6, 2010* (September 30, 2010), 2, http://www.sec.gov/news/studies/2010/marketevents-report.pdf.

[54] Ibid, 3-4.

Since then, the SEC's initial report has been highly disputed, not only by outside researchers but also seemingly by other parts of the government. In April 2015, the U.S. Justice Department issued an indictment against a London-based trader for market manipulation using high-frequency trading algorithms (behavioral hacking).[55] The head of enforcement at the U.S. Commodity Futures Trading Commission stated, "[The trader's] conduct was at least significantly responsible for the order imbalance that in turn was one of the conditions that led to the flash crash."[56] Independent research points to what was likely a confluence of multiple factors, which is common for normal accidents.[57]

What appears clear across multiple analyses of the May 2010 incident is that automated stock trades and high-frequency trading algorithms at the very least played a role in exacerbating the crash. This may have been due, in part, to unanticipated interactions between adversarial trading algorithms. It is also possible that behavioral hacking of the algorithms was a factor.

Adversarial risk exacerbates the problem of normal accidents

The 2010 flash crash demonstrates the risk of highly complex autonomous systems interacting in competitive environments. The risk of unanticipated interactions is increased, since competitors are not likely to share their algorithms with one another. In fact, when the behavior of the autonomous system can be predicted, it is susceptible to behavioral hacking by adversaries.

Moreover, in this environment, information is incomplete and time is precious. While humans maintain supervisory control over the stock market in principle, the speed of interactions means that the potential damage high-frequency trading algorithms can cause before humans take corrective action may be quite high. Human control is akin to that of an inattentive human driver on an autonomous car speeding down the highway—steering wheel or no, the driver is a *de facto* passenger along for the ride.

In response to the 2010 flash crash, federal regulators have pursued a number of measures to prevent such incidents in the future, including "circuit breakers" that would halt trading if stock prices dropped too quickly.[58] However, mini-flash crashes have continued to be reported.[59] An average day

---

[55] Douwe Miedema and Sarah N. Lynch, "UK Speed Trader Arrested over Role in 2010 'Flash Crash'," *Reuters*, April 21, 2015, http://www.reuters.com/article/2015/04/21/us-usa-security-fraud-idUSKBN0NC21220150421.

[56] Ibid.

[57] Torben G. Andersen and Oleg Bondarenko, "VPIN and the Flash Crash," *Journal of Financial Markets* 17 (May 8, 2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1881731; David Easley, Marcos Lopez de Prado, and Maureen O'Hara, "The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading," *The Journal of Portfolio Management*, 37, no. 2 (Winter 2011), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1695041; and Wes Bethel, David Leinweber, Oliver Ruebel, and Kesheng Wu, "Federal Market Information Technology in the Post Flash Crash Era: Roles for Supercomputing," *Proceedings of the Fourth Workshop on High Performance Computational Finance* (September 25, 2011), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1939522.

[58] Bob Pisani, "'Flash Crash' 5 years later: What have we learned?" *CNBC*, May 5, 2015, http://www.cnbc.com/id/102651458.

[59] Eric Garcia, "Two Mini-flash Crashes Rock Stock Market Tuesday," MarketWatch, November 25, 2014, http://www.marketwatch.com/story/two-mini-flash-crashes-rock-stock-market-2014-11-25?dist=tcountdown.

of trading sees a handful of circuit breakers tripped due to rapid price drops. On one day in August 2015, over 1,000 circuit breakers were tripped across multiple exchanges.

The problem of flash crashes in stock trading points to the dangers of competitive autonomous systems interacting at high speed in adversarial environments. One important difference between stock markets and military environments, however, is that stock markets have external regulators who can enforce stabilizing measures like circuit breakers. In war, there is no arbiter to call "time out" if conflicts begin to spiral out of control. Militaries will have to institute their own fail-safe measures to ensure that autonomous systems do not lead to catastrophic outcomes.

VIII. The Human as Fail-Safe

Much of the discourse on autonomous weapons to date has focused on whether their use would be legal and ethical, but an equally important question is whether they could be used safely. Even if they could be used in a manner that is lawful and ethical under most operating conditions, it is conceivable that they could be quite dangerous. The consequences of a failure with some types of autonomous weapons could be catastrophic. Autonomous weapons, like other complex systems, are susceptible to failure. While better design, testing, and operator training can decrease the likelihood of these failures, they cannot be eliminated entirely. Some failures will inevitably occur. It is crucial that when failures occur, the systems fail *safe.*[60] This entails minimizing the potential damage resulting from a failure.

Consider, for example, the Patriot fratricides. Humans were in the loop for both engagements, and yet the fratricides still occurred. Humans, after all, make mistakes as well, and human errors no doubt contributed to the fratricides. However, human involvement did prevent individual incidents from cascading into mass fratricide. Human operators were able to regain control of the weapon system quickly once it became apparent that it was not performing appropriately, and halt its operation.

Imagine an alternate scenario in which the Patriot had been operating fully autonomously, with no ability for human operators to observe its functioning and halt its operation. The cause of the F-18 fratricide was rare enough that it likely would not have happened again. It is entirely possible that the conditions that led to the Tornado shoot-down, on the other hand—misidentification of the aircraft as an anti-radiation missile and IFF failure—would have been replicated again, given enough Patriot-aircraft interactions. In fact, in a near-miss incident a day after the Tornado shoot-down, an Air Force F-16 pilot fired on a Patriot radar battery that had locked onto his aircraft. The F-16 destroyed the Patriot's radar, but no one was killed.[61] In training, Patriot batteries had "fired" on friendly aircraft repeatedly.[62] A fully autonomous version of the Patriot with no human to halt the system's operation could have resulted in far more fratricides.

While the Patriot fratricides were tragic, they did not change the course of the war. By contrast, a fully autonomous weapon system that began engaging friendlies could run rampant, committing mass fratricide until it exhausted its ammunition. Adversaries would have strong incentives to hack

[60] In some cases, "fail safe" will mean that the system itself automatically reverts to a safer mode of operation, such as when a nuclear reactor automatically scrams, dropping control rods into the reactor core. These safeties are important, but ultimately—as the Three Mile Island and Fukushima incidents demonstrate—have their limitations. When humans function as a fail safe, such as in commercial airliners, safe operation may not result instantly once the human assumes control. Some active control on the part of the human operator may be required to get the system to a state of safe operation.

[61] David Axe, "That Time an Air Force F-16 and an Army Missile Battery Fought Each Other Fighter: Pilots Feared Flawed Air-defense System," War is Boring, July 5, 2014, https://medium.com/war-is-boring/that-time-an-air-force-f-16-and-an-army-missile-battery-fought-each-other-bb89d7d03b7d.

[62] For an overview of some of the earlier incidents identified in training, see Pamela Hess, "Feature: The Patriot's Fratricide Record," *UPI*, April 24, 2003, http://www.upi.com/Business_News/Security-Industry/2003/04/24/Feature-The-Patriots-fratricide-record/63991051224638/.

such systems, either directly through malware or via behavioral hacking, to turn them on friendly forces.

Humans exhibit some inherent resiliency against hacking. They can ignore orders if they don't make sense. They can use common sense to adapt to the situation at hand. An order coming across the radio to "attack friendly forces" might easily be ignored as a not-very-clever enemy ruse. Target coordinates that turned out to be one's own base could be dismissed as an error. Yet autonomous systems lack the flexibility to ignore orders or consider the broader context. The Patriot system was not aware that it had shot down a friendly aircraft; it lacks the sensors and cognitive ability to even make that assessment.

Humans vs. automation is a false choice

Automation is good at many things—precision, reliability, and speed among them. But autonomous systems are brittle. The lack the flexibility humans have to adapt to novel situations. What would an autonomous system have done if it was in the same situation Stanislav Petrov found himself on September 26, 1983? Whatever it was programmed to do.

Are we doomed, then, to choose between the brittleness of automation or human cognitive weaknesses? Automation already has increased precision in war with the advent of precision-guided munitions, dramatically reducing civilian casualties.[63] AIs already perform equally as well as or better than humans at visual object recognition most of the time, and they are improving.[64] Is the price for gaining these advantages accepting the rare instances when autonomous systems lead to accidents, even potentially catastrophic ones?

No—humans vs. machines is a false choice. The best systems will combine human and machine intelligence to create hybrid cognitive architectures that leverage the advantages of each. As an example of the future of cognition, look no further than one of the most high-profile areas in which AIs have bested humans: chess.

The best chess players in the world are human-machine teams

In 1997, world chess champion Gary Kasparov lost to IBM's Deep Blue, cementing the reality that humans are no longer the best chess players in the world. But neither, as it turns out, are machines. A year later, Kasparov founded the field of "advanced chess," or centaur chess, in which humans and AIs cooperate on the same team. By leveraging the advantages of human *and* machine, centaur chess results in a more perfect game, better than humans or AIs alone. The AIs can analyze possible moves and identify vulnerabilities or opportunities the human player might have missed, resulting in

[63] Michael Horowitz and Paul Scharre, "Do Killer Robots Save Lives?" *Politico,* November 19, 2014, http://www.politico.com/magazine/story/2014/11/killer-robots-save-lives-113010.

[64] Stuart Russell, "Artificial Intelligence: Implications for Autonomous Weapons" (United Nations Convention on Certain Conventional Weapons, Geneva, Switzerland, April 13, 2015), http://www.cs.berkeley.edu/~russell/talks/russell-ccw15-autonomy.pptx.

blunder-free games. The human player can manage strategy, prune AI searches to focus on the most promising areas, and manage differences between multiple AIs.[65] The chess AI, or multiple AIs, gives feedback to the human player, who then decides what move to make.[66]

Similarly, combining human and machine cognition for engagement decisions could yield the precision and reliability of automation without sacrificing the robustness and flexibility that humans bring.

[65] Tyler Cowen, "What are Humans Still Good for? The Turning Point in Freestyle Chess may be Approaching," Marginal Revolution, November 5, 2013. http://marginalrevolution.com/marginalrevolution/2013/11/what-are-humans-still-good-for-the-turning-point-in-freestyle-chess-may-be-approaching.html.

[66] This dynamic changes in timed games where 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 30-60 seconds to make a move. Over time, as computers advance, one would anticipate this time horizon to expand until humans no longer add any value regardless of how much time is allowed. Cowen, ibid. This situation is analogous to the role human-supervised autonomous weapons play today in defending against saturation attacks from missiles and rockets, where the speed of engagements could easily overwhelm human operators' ability to respond quickly enough.

IX. Centaur Warfighting

Human-machine teaming is a better approach than using humans or autonomous systems alone, bringing to bear the unique advantages of each. Understanding how human-machine teaming, or "centaur warfighting," might work in engagement decisions requires first disaggregating the different roles a human operator performs today with respect to selecting and engaging enemy targets.

In today's semi-autonomous weapon systems, humans currently perform three kinds of roles with respect to target selection and engagement. In some cases, human operators perform multiple roles simultaneously.

- **The human as essential operator:** The weapon system cannot accurately and effectively complete engagements without the human operator.

- **The human as moral agent:** The human operator makes value-based judgments about whether the use of force is appropriate—for example, whether the military necessity of destroying a particular target in a particular situation outweighs the potential collateral damage.

- **The human as fail-safe:** The human operator has the ability to intervene and alter or halt the weapon system's operation if the weapon begins to fail or if circumstances change such that the engagement is no longer appropriate.

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 infra-red sensor in the nose of the weapon to detect the target. [One of the aircraft] 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.[67]

[67] Mike Pietrucha, "Why the Next Fighter will be Manned, and the One After That," War on the Rocks, August 5, 2015, http://warontherocks.com/2015/08/why-the-next-fighter-will-be-manned-and-the-one-after-that/.

In this example, the pilots were performing all three roles simultaneously. In manually guiding the air-to-ground weapon they were acting as essential operators. Without their guidance, the weapon would not have been accurate or effective. They also were 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. Finally, 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.

In a different scenario, human operators might perform only 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 as moral agents and fail-safes, even if they were no longer essential operators once they launched the weapon.[68]

Other types of automation in non-military settings disaggregate these functions in various ways. A person kept on medical life support has machines performing the essential task of keeping him or her alive, but humans are making the moral judgment whether to continue life support. 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.

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 the weapon system's operation may result in far greater accuracy, precision, and reliability than relying on a human operator. Just as autonomous systems can land airplanes, manage subway repair schedules, play chess, answer trivia questions, and arrive at complex medical diagnoses more accurately than humans, they also may be capable of performing many tasks in war better than humans.[69] Automating the human's role as moral agent or fail-safe, however, may be far harder. Humans have moral and legal judgment, responsibility, and accountability, making their role as moral agents important for many tasks in war. Humans also have great value as fail-safes, with the ability to flexibly respond to a range of unplanned scenarios.

[68] This assumes that they have sufficient time to perform these roles as moral agent and fail-safe, which will depend on the specific situation. The same pressures that drove the desire to automate the essential operation of the weapon system could complicate the human's ability to act as moral agent or fail-safe.

[69] Hal Hodson, "The AI Boss that Deploys Hong Kong's Subway Engineers," *New Scientist*, July 4, 2014, http://www.newscientist.com/article/mg22329764.000-the-ai-boss-that-deploys-hong-kongs-subway-engineers.html#.VRB7jELVt0c; and "What is Watson?," IBM, http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html.

Human-machine teaming in engagement decisions

It is possible to design systems that incorporate both automation *and* human decision-making, using automation to perform essential tasks with greater precision and accuracy while retaining humans in the roles of moral agents and fail-safes.

The U.S. counter-rocket, artillery, and mortar (C-RAM) system is an example of this approach, automating much of the engagement, resulting in more precise and accurate engagements, while keeping 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, using a network of radars to automatically identify and track incoming rounds. Because the C-RAM is frequently used at U.S. air bases where there are friendly aircraft in the sky, the system autonomously creates a "Do Not Engage Sector" around friendly aircraft to prevent fratricide.[70] The result is a highly automated system that, in theory, would be 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.[71]

Thus, the C-RAM employs overlapping safeties, both automated and human. The autonomous 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.[72]

In principle, an approach along the lines of C-RAM's blended use of automation and human decision-making is optimal, leveraging the advantages of each. This allows militaries to add automation to increase precision and accuracy without giving up the role of the human as moral agent and fail-safe. From the perspective of normal accident theory, the human in the loop creates a buffer in the system, reducing the degree of the system's coupling. The human operator is analogous to a "circuit breaker" in financial markets. Instead of one failure cascading to many unintended engagements, the human in the loop decouples each engagement from the others, allowing the human to react and adjust as needed between engagements.

---

[70] Mike Van Rassen, "Counter-Rocket, Artillery, Mortar (C-RAM)," Program Executive Office Missiles and Space, Slide 28, http://www.msl.army.mil/Documents/Briefings/C-RAM/C-RAM%20Program%20Overview.pdf.

[71] Sam Wallace, "The Proposed Ban on Offensive Autonomous Weapons is Unrealistic and Dangerous," Kurzweilai, August 5, 2015, http://www.kurzweilai.net/the-proposed-ban-on-offensive-autonomous-weapons-is-unrealistic-and-dangerous.

[72] The C-RAM was designed post-2003, after the Patriot fratricides. One can understand why this dual-safety approach was desirable, given the Patriot's record in OIF.

In order for the human operators to actually perform the roles of moral agent and fail-safe, the operators must be trained for and supported by a culture of active participation in the weapon system's operation. The type of "unwarranted trust" in automation that led to the Patriot fratricides would result in a human in the loop in name only. Training that requires human operators to exercise judgment and a culture that emphasizes human responsibility are essential to ensuring that the human's role remains meaningful.

Human-supervised autonomous weapons

Keeping a human in the loop, even if only as a fail-safe, has many advantages, but there may be situations where keeping a human in the loop is simply not feasible. This could be because the speed of engagements exceeds the ability of human operators to respond. While humans remain in the loop for C-RAM, at least 30 countries, including the United States, employ automated defensive systems similar to C-RAM but with modes that shift to a human-supervised, on the loop control type.[73] Once these modes are activated, human operators can observe the weapon system's operation and can intervene if necessary, but the weapon will not wait for human authorization before firing.

These weapons entail a higher degree of risk than semi-autonomous (human in the loop) systems, such as C-RAM. The damage potential of the system depends on how quickly human operators can identify that the system is failing and take corrective action. In some situations, multiple unintended engagements could occur.

Another crucially important factor in human-supervised control is the reliability of communications between the human controller and the autonomous system and what the system will do if it loses communications. If communications are lost with the human operator, will the system halt engagements (fail-safe) or continue engaging targets, now as a fully autonomous weapon (fail-deadly)?[74]

The degree of physical access to the system is another important factor in assessing risk. Today's human-supervised autonomous weapons are used onboard human-occupied ships, bases, or ground vehicles where human operators have physical access to the system. This is important in two critical ways. First, communications are hardwired and do not depend on wireless transmission. Second, in the event of a failure, human operators can physically disable the system to prevent further engagements. If the human operators were supervising the system remotely, however, their ability to exercise effective control would be through software, creating another vulnerability. A common mode failure (such as enemy hacking) that caused unintended engagements *and* negated the ability of human operators to retake control of the system could lead to significant damage.

[73] Paul Scharre and Michael Horowitz, "An Introduction to Autonomy in Weapon Systems," Center for a New American Security, February 2015, Appendix B.

[74] From an operational perspective, of course, there may be situations where a "fail deadly" model is operationally preferable and entails less risk overall to friendly forces. This should be carefully assessed, however, based on an appreciation of the various risks and benefits, factoring in the potential for accidents, enemy hacking, and unanticipated situations.

Fully autonomous weapons

Human supervision is only possible if the weapon system has reliable, real-time or near-real-time communications with human operators. This may not always be the case, however. Communications are challenging in some environments, such as underwater, and adversaries will seek to jam or disrupt communications. This may drive militaries to consider fully autonomous weapons (or supervised autonomous weapons that "fail deadly" if they lose communications).[75] Full autonomy would allow uninhabited vehicles to carry out engagements against emergent targets of opportunity without specific human authorization. Alternatively, uninhabited systems could be designed to "fail dangerous" in the event of a loss of communications—they would only strike pre-planned human-authorized targets offensively, but could engage in limited self-defense to prevent the vehicle's destruction.

Thus, rules of engagement for uninhabited systems operating without communications to human controllers could be grouped into three broad categories:

- **Fail-safe**: In the event of communications loss, the uninhabited vehicle only engages targets that have been pre-authorized by human controllers, similar to homing missiles, torpedoes, or cruise missiles today. The vehicle cannot use lethal force to defend itself against emergent threats, only jamming or other non-lethal measures. (Semi-autonomous operation for offensive and defensive actions.)

- **Fail-dangerous**: In the event of a communications loss, the uninhabited vehicle only offensively engages targets that have been pre-authorized by human controllers. The vehicle is authorized to use limited, proportional lethal force to defend itself from attack. (Semi-autonomous operation for offensive action; fully autonomous operation for limited self-defense.)

- **Fail-deadly**: In the event of a communications loss, the uninhabited vehicle can engage emergent targets of opportunity that have not been specifically approved by human operators. It can also use lethal force to defend itself. (Fully autonomous operation for offensive and defensive actions.)

For each of these rules of engagement options, militaries will want to think hard about not only the military value of this degree of autonomy if it works correctly, but the potential consequences if the autonomy fails or the system is hacked.

[75] For more on issues relating to autonomy in the maritime environment, see United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters*, no. 4 (2015), http://www.unidir.org/files/publications/pdfs/testing-the-waters-en-634.pdf.

Weighing the operational risk and value of autonomous weapons

Perrow recommends weighing the net catastrophic potential of a complex, tightly coupled system against the cost of an alternative. For some systems, the risk may be so great that forgoing them entirely may be the best option.[76]

The bulk of this paper has focused on the risk associated with using autonomous weapons. However, their value—or, alternatively, the risk of not using them— also should be considered.

The operational value of defensive human-supervised autonomous weapons is clear. Saturation attacks from rockets and missiles could overwhelm human operators, a reality that has led over 30 nations to acquire air, rocket, and missile defense systems with human-supervised autonomous modes. Future advances in autonomy and swarming are likely to only exacerbate this trend. Future threats could appear in greater numbers, maneuver more quickly, and coordinate their attacks. Presently, these human-supervised autonomous weapons are used in fairly narrow circumstances, however. They are used to defend human-occupied bases or vehicles. They do not target people, only objects or enemy vehicles. And human operators have physical access to the weapon system to disable it in the event of a failure.

A faster pace of operations, perhaps driven by an adversary's autonomous weapons, could expand the number of situations in which human-supervised autonomous weapons would be desired. Just as autonomy is needed to successfully defend against saturation attacks from missiles and rockets today, it similarly might be needed in other future situations where human reaction times are too slow to be successful. However, provided there is adequate communications, keeping a human on the loop to supervise the system's operation would still be possible and desirable, just as it is for defensive systems today. In fact, for many applications, it may still be possible to keep a human in the loop in a semi-autonomous mode of operation without sacrificing much in the way of a time delay. Humans would not need to physically maneuver the weapon—after all, homing missiles and torpedoes are "fire and forget" today—but only remain in charge of authorizing targets for engagement.

The operational value of fully autonomous weapons is less clear. Fully autonomous weapons might be desired if communications were entirely lost with human controllers. The use of force could be offensive, to engage emerging targets of opportunity that have not been pre-authorized by human controllers, or defensive to defend uninhabited vehicles.[77] While giving uninhabited systems some limited ability to defend themselves in the event of a communications loss seems potentially valuable, it is not clear how necessary offensive fully autonomous weapons would be for effective operations in contested areas.[78]

---

[76] Perrow, *Normal Accidents,* 342-352.

[77] By definition, a weapon that carried out attacks against fixed targets that had been previously authorized by human controllers would be a semi-autonomous, not a fully autonomous, weapon.

[78] One conceivable argument for fully autonomous weapons might be to deliberately sever the communications link with human controllers to minimize vulnerability to hacking. It is worth pointing out that this would eliminate one potential vector for hackers to gain access to the system, but would not render the system hacker-proof. Any system with a computer is susceptible to malware, which could still be introduced through other means, such as when the system is connected during maintenance. USB flash drives, for example, are notorious for spreading malware across military computer systems. While severing communications

Communications in contested areas is not an all-or-nothing proposition. Capable militaries will be able to employ jam-resistant communications, although they will be limited in bandwidth and range.[79] This could allow a human in a nearby vehicle to remotely remain in the loop to authorize engagements. In this concept of operations, human-machine teaming occurs both physically and cognitively. Physically, both uninhabited and human-inhabited vehicles would operate forward in the battlespace. Uninhabited vehicles would likely operate at the vanguard of the formation, with human controllers quarterbacking the fight from nearby vehicles further removed from enemy threats. Cognitive tasks are likewise shared by humans and autonomous systems. Many routine tasks could be automated, but key decisions, such as selecting and engaging targets, would be relayed to nearby human controllers for approval. Thus, militaries could exploit many of the advantages of uninhabited and autonomous systems, including their ability to take greater risk, in contested environments while still keeping a human in the loop for engagement decisions.

Again, a key question would be whether the systems would be designed to fail safe, dangerous, or deadly in the event that communications failed. Even if militaries did not field offensive fully autonomous weapons, a limited "fail-dangerous" posture that allowed uninhabited systems to autonomously exercise limited, proportional lethal force to defend themselves from attack might be warranted. While technically such engagements would be fully autonomous, they would be of a limited nature.

Anti-personnel autonomous weapons

Anti-personnel autonomous weapons deserve special mention because of their increased potential risk to civilians. The military utility of a fully autonomous anti-personnel weapon is questionable, at best.[80] Unlike defending against missiles, which can travel up to hypersonic speeds, the additional time it would take for a defender to authorize engagements would be marginal compared to the speed at which attackers can run on foot. Overwhelming defensive positions through waves of human attackers has not been an effective tactic since the invention of the machine gun. Offensive anti-personnel attacks in communications-denied environments also would likely not be necessary, as communications-contested environments are likely to be those where targeting military systems—radars, missile launchers, airfields, aircraft, etc.—is most useful. Some limited automated defensive measures might be necessary to prevent tampering if uninhabited vehicles were unattended and out of communications, although non-lethal measures would likely be most appropriate.

---

would take away one potential vector for attacks, doing so would come at a high cost: forgoing any ability to retake control of the system or retask it once it is launched, even if began performing inappropriately. Computer security, protected communications, and a process for authenticating valid authorization for commands are essential to all networked, computerized military systems, autonomous or not. However, opting for fully autonomous weapons because of fears about hacking would be a strange choice. While the probability of an adversary gaining access might be somewhat reduced, the potential consequences if an adversary were to gain access could be far more severe. The net balance of risk is likely to favor opting for increased opportunity for human control, where possible.

[79] Scharre, "Yes, Unmanned Combat Aircraft are the Future."

[80] A weapon that targeted specific individuals that had previously been selected by humans would, by definition, be a semi-autonomous weapon since a human would be choosing the target.

It is beyond the scope of this paper to examine all possible situations where supervised-autonomous or fully autonomous weapons may have some operational utility. Other operational scenarios are also possible, some of which may be more or less likely. These examples are included merely to illustrate some of the potential factors that affect military utility. What is important is that if militaries consider employing autonomous weapons, they include an assessment of their risks, given their higher damage potential relative to semi-autonomous weapons. Given the inevitability of failures over a long enough period of operational use, the military necessity of autonomous weapons must be quite high to warrant accepting the risk of their employment.

X. Assessing and Managing the Risks of Autonomous Weapons

While careful risk assessments of autonomous weapons are essential, policymakers, independent experts, and military professionals should be skeptical of their confidence level in any estimation of the risk of employing an autonomous weapon. Understanding risks associated with low probability, high consequence events is notoriously difficult, and militaries' track records in managing risks of this type are mixed at best.

Accurately assessing the risk of low-probability accidents is very difficult

Simply accurately estimating the risk of a low probability accident can be exceedingly challenging. In an appendix to the official report on the *Challenger* accident, Nobel prize–winning physicist Richard Feynman noted the wide disparity of views within NASA regarding the probability of accidents:

> It appears that there are enormous differences of opinion as to the probability of a failure with loss of vehicle and of human life. The estimates range from roughly 1 in 100 to 1 in 100,000. The higher figures come from the working engineers, and the very low figures from management. What are the causes and consequences of this lack of agreement? Since 1 part in 100,000 would imply that one could put a Shuttle up each day for 300 years expecting to lose only one, we could properly ask "What is the cause of management's fantastic faith in the machinery?"[81]

While Feynman's observation was about space shuttle operations, some conclusions can be drawn about the difficulty of assessing risk in complex systems more generally.

First, actually quantifying the likelihood of accidents is very challenging. The figures Feynman cites are judgments given by people about how likely they think an accident is. There is disagreement precisely because it is difficult to objectively quantity the risk. More data and testing can help people make more accurate judgments, but there is no straightforward formula for calculating the likelihood of an accident with a sufficiently complex system.

Second, social and organizational factors clearly influence the judgments people make about risk. Regardless of whether the actual risk of shuttle accidents was closer to the assessments of the engineers or managers, the fact that there is a systematic disparity suggests other organizational and social factors at work that bias these estimations.[82] (The loss of *Columbia* less than 100 missions after *Challenger* suggests that the engineers' estimates were closer to reality, however.)

---

[81] Richard P. Feynman, "Volume 2: Appendix F – Personal Observations on Reliability of Shuttle," http://history.nasa.gov/rogersrep/v2appf.htm.

[82] It is possible that this difference in risk assessment could be due to an information asymmetry, that the engineers simply knew more about the systems than the managers, but in a functional bureaucracy, this risk should be translated accurately up the chain of command. The fact that it was not suggests some social or organizational component that either distorted the risk calculus or prevented accurate information flow about risk.

Finally, the low probabilities cited in all estimates highlight the challenge in understanding low probability, high consequence risks. Feynman points out what a 1 in 100,000 accident rate would mean in the case of shuttle operations, but the real-world implications of a 1 in 100,000 vs. a 1 in 100 accident rate may not be intuitive to most people who are not used to estimating low probability events.

For example, if policymakers were told an autonomous weapon had a 1 in 10 chance of fratricide, they might reasonably avoid deploying such a system. However, if they were told that a system had a 1 in 10,000 chance of fratricide (99.99% safety rate), verified by testing, they might reasonably conclude that such a system was fairly safe. The odds of an accident would seem low. But if the number of potential interactions with friendly forces in a combat environment numbered in the "millions," as the Defense Science Board noted was the case with the Patriot, the actual number of fratricides could still be in the hundreds in a major military campaign, enough to have significant operational impact. Yet for those not used to assessing low probability, high consequence risk, a 1 in 10,000 risk might seem quite safe.[83]

High-reliability organizations

One potential response to normal accidents is high-reliability organizations, organizations that exhibit certain characteristics that allow them to routinely operate high-risk systems with low accident rates.[84] High-reliability organizations can be found across a range of industries but exhibit certain common characteristics. These include not only high-trained individuals, but also a collective mindfulness of the risk of failure and a continued commitment to learn from near-accidents and improve safety.[85]

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. One example is the U.S. Navy's submarine community. 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:

> 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

[83] The important distinction is that the risk of an accident actually occurring depends on the risk of accident in any given instance times the number of exposures to that risk over a given period.
[84] High-reliability organizations do not necessarily have a zero accident rate, however.
[85] Karl E. Weick and Kathleen M. Sutcliffe, *Managing the Unexpected: Sustained Performance in a Complex World, 3rd edition,(CITY?:* Jossey-Bass, 2015).

> 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.[86]

The Navy's SUBSAFE program has been able to substantially reduce the risks associated with an inherently dangerous task: operating military submarines. However, whether this model could be applied to autonomous weapons is questionable. The U.S. Navy submarine community is a very specific military community. Other elements of the U.S. military do not necessarily exhibit the same characteristics as the SUBSAFE program, nor do other nations' submarine communities. The accident rate with Soviet/Russian submarines is far higher, for example. Furthermore, it should be noted that while SUBSAFE has reduced the risks associated with submarine operations, it cannot eliminate them entirely. Attempting to apply the same rigorous high-reliability standards used in the SUBSAFE program to all autonomous weapons across multiple military communities and many countries would be effectively impossible.

Nuclear weapons safety and near-miss accidents

Nuclear weapons are an instructive example in the challenges in managing the risks associated with extremely dangerous weapons. Nuclear weapons individually hold the potential for significant mass destruction. Collectively, a nuclear exchange could destroy human civilization. But outside of testing they have not been used, intentionally or accidentally, since 1945.

The safety track record of nuclear weapons is less than inspiring, however.[87] In addition to the Stanislav Petrov incident in 1983, there have been a number of nuclear near incidents that could have had catastrophic consequences. A 2014 report by Chatham House highlighted 13 such incidents from 1962-2002, some of which could have resulted in an individual weapon's use and others which potentially could have resulted in a nuclear exchange between superpowers. These include:

- 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.[88]
- Less than a year later in 1980, a faulty computer chip led to a similar false alarm at NORAD. This incident led U.S. military commanders to notify National Security Advisor Zbigniew Brzezinski that 2,200 Soviet missiles were inboard to the United States. Brzezinski was about to inform President Jimmy Carter before NORAD realized the alarm was false.[89]

---

[86] Paul E. Sullivan, "Statement before the House Science Committee on the SUBSAFE Program," October 29, 2003, http://www.navy.mil/navydata/testimony/safety/sullivan031029.txt.
[87] For an in-depth analysis of nuclear weapons from the perspective of normal accidents, see Scott D. Sagan, "The Limits of Safety: Organizations, Accidents, and Nuclear Weapons."
[88] Lewis et al., "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy," 12-13.
[89] Lewis et al., ibid, 13.

- 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, it could have been 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. The Russian nuclear briefcase was brought to President Boris Yeltsin, who discussed a response with senior Russian military commanders before the missile was identified as harmless.[90]

In addition to these incidents are safety lapses that might not have risked a nuclear accident but are troubling nonetheless. For example, in 2007 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 security lapses in the U.S. nuclear community, with Air Force leaders citing an "erosion" of adherence to appropriate weapons handling safety standards.[91]

These incidents do not inspire confidence. Safety is challenging enough with nuclear weapons. Autonomous weapons would potentially be more difficult in a number of ways. They could be proliferated more widely to actors less capable or those less interested in safe operation, including rogue regimes such as North Korea. The increased tempo of operations enabled by adversary autonomous weapons could exacerbate adversarial risks, limiting the ability of military commanders and political leaders to analyze the situation before reacting.

While autonomous weapons almost certainly would have a lower inherent hazard than nuclear weapons, this fact itself could make risk mitigation more challenging. Nuclear weapons clearly have massive destructive potential, but militaries may perceive autonomous weapons as safe— and therefore as not requiring special treatment—because of the precision inherent in automation. Yet the consequences of accidents with autonomous weapons could still be quite severe. If militaries cannot reliably institute safety procedures to control and account for nuclear weapons, their ability to safely control autonomous weapons is far less certain.

---

[90] Lewis, "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy," 16-17.

[91] Defense Science Board Permanent Task Force on Nuclear Weapons Surety, "Report on the Unauthorized Movement of Nuclear Weapons," February 2008, http://web.archive.org/web/20110509185852/http://www.nti.org/e_research/source_docs/us/department_defense/reports/11.pdf; and Richard Newton, "Press Briefing with Maj. Gen. Newton from the Pentagon, Arlington, Va.," October 19, 2007, http://web.archive.org/web/20071023092652/http://www.defenselink.mil/transcripts/transcript.aspx?transcriptid=4067.

XI. Conclusion

War is a hazardous endeavor. Militaries must balance various kinds of risk—risk to their forces, the mission, their citizens, innocent civilians, and possibly to the state itself. Military personnel risk their own lives in combat and, in some wars, the state's very survival may be at stake. States also balance risk among strategic objectives: deterrence, defense, and crisis stability. Military forces must be ready to respond at a moment's notice to provocation, for example, but not on such a hair-trigger that they create a crisis or cause one to escalate unnecessarily. Militaries will come to different conclusions on how to weigh these risks, and their strategic position vis-a-vis external threats is a major factor. Military underdogs or those facing threats to the state's very existence may, quite logically, be willing to take bigger operational risks to achieve their aims. In some situations, states engage in deliberately risky behavior as a tactic of brinksmanship, to deter or coerce others.

No country has stated that they plan to build fully autonomous weapons, but few countries have renounced them either. Over 90 countries and non-state groups already have uninhabited aircraft, or drones. Today, these drones are largely remotely controlled, but over time next-generation versions will incorporate greater autonomy. When there is sufficient communications to keep a human in the loop, there is great value in doing so. Humans can act as a fail-safe and are flexible enough to respond to a wide array of situations. However, when communications are denied, autonomous weapons would allow engagements against targets of opportunity and allow uninhabited vehicles to defend themselves from attack.

As technology advances, militaries must carefully consider the risks of employing autonomous weapons. Much of the debate on autonomous weapons focuses on legal, moral, or ethical issues. Autonomous weapons also raise critically important issues of controllability and safety, however. It is possible to envision weapons that would perform lawfully most of the time but that in the event of a failure could lead to catastrophe. A loss of control with an autonomous weapon could lead to mass fratricide or civilian casualties, or cause a crisis to spiral out of control. Over a long enough period of operational use, some failures are inevitable. Employing autonomous weapons would mean accepting the consequences of these inevitable failures.

Greater transparency is needed among states on how they will approach autonomy in weapon systems. Few states have issued clear national policies on the use of autonomy in weapons. Given the potential for dangerous interactions between autonomous systems, a common set of international expectations is critical. 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.

While much of the discourse on autonomous weapons to date focuses on the very important questions of humanitarian impact to civilians, autonomous weapons also raise important questions of strategic stability. This is an important aspect of autonomous weapons that deserves further

consideration. States have a long history of cooperating to regulate, ban, or develop common norms and expectations for a variety of weapons that were seen as destabilizing and dangerous—nuclear weapons, space-based and counter-space weapons, anti-ballistic missile weapons, and intermediate range nuclear-capable missiles, to name a few. Continued international dialogue to develop common norms, or "rules of the road," for the use of autonomy in weapons is necessary to help manage the potential strategic risks of autonomous weapons and avoid dangerous outcomes.