CNA
ANALYSIS & SOLUTIONS

September 2019

AI Safety: An Action Plan for the Navy

Larry Lewis

DOP-2019-U-021957-Final

Abstract

In light of the Navy's stated commitment to using AI, and given the strategic importance of AI safety, we provide the Navy with a first step towards a comprehensive approach to safety. We use a risk management approach to frame our treatment of AI safety risks: identifying risks, analyzing them, and suggesting concrete actions for the Navy to begin addressing them. The first type of safety risk, being technical in nature, will require a collaborative effort with industry and academia to address. The second type of risk, associated with specific military missions, can be addressed in a combination of military experimentation, research, and concept development to find ways to promote effectiveness along with safety. For each types of risk, we use examples to show concrete ways of managing and reducing the risk of AI applications. We then discuss institutional changes that would help promote safety in the Navy's AI efforts.

This document contains the best opinion of CNA at the time of issue.

It does not necessarily represent the opinion of the sponsor

Distribution

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

Public Release.

9/20/2019

This work was performed under Federal Government Contract No. N00014-16-D-5003.

Cover image credit: : From Artificial Intelligence, Thinking Machines and the Future of Humanity (Gerd Leonhard)

Approved by: **September 2019**

Mr. Mark B. Geis
Executive Vice President, FFRDC
CNA/CNA

Request additional copies of this document through inquiries@cna.org.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
09-2019

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
(U) AI Safety: An Action Plan for the Navy

5a. CONTRACT NUMBER
N00014-16-D-5003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0605154N

6. AUTHOR(S)
Larry Lewis

5d. PROJECT NUMBER
R0148

5e. TASK NUMBER
D665.00

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Naval Analyses
4825 Mark Center Drive
Alexandria, VA 22311-1850

8. PERFORMING ORGANIZATION REPORT NUMBER
DOP-2019-U-021957-Final

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Chief of Naval Operations
(OPNAV N81)
Navy Department Pentagon
Washington, DC 20350

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
In light of the Navy's stated commitment to using AI, and given the strategic importance of AI safety, we provide the Navy with a first step towards a comprehensive approach to safety. We use a risk management approach to frame our treatment of AI safety risks: identifying risks, analyzing them, and suggesting concrete actions for the Navy to begin addressing them. The first type of safety risk, being technical in nature, will require a collaborative effort with industry and academia to address. The second type of risk, associated with specific military missions, can be addressed in a combination of military experimentation, research, and concept development to find ways to promote effectiveness along with safety. For each types of risk, we use examples to show concrete ways of managing and reducing the risk of AI applications. We then discuss institutional changes that would help promote safety in the Navy's AI efforts.

15. SUBJECT TERMS
AI, safety, fratricide, civilian casualties, human machine teaming, autonomy

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	SAR	50

19a. NAME OF RESPONSIBLE PERSON
Knowledge Center/Robert Richards

19b. TELEPHONE NUMBER *(include area code)*
703-824-2123

Executive Summary

History is replete with examples of technology being leveraged for military advantage. The chariot, the crossbow, gunpowder, and nuclear weapons all brought profound, disruptive effects to the battlefield. Many observers believe Artificial Intelligence (AI) will have the same or greater effects on warfare. It is not surprising, then, that nations are planning to take advantage of this technology. In fact, increasingly more countries see advances in AI as central to their national security strategies.

As the US, China, Russia, and other nations pursue AI, plans and efforts to leverage it for military applications have encountered significant concerns. For example, Google has announced that it will no longer support the Department of Defense's (DOD's) Project Maven, and the UN Forum on Certain Conventional Weapons is holding talks, during which some parties have urged a pre-emptive ban on lethal autonomous weapon systems, one possible application of AI to warfare. When the media covers military applications of AI, it often use scenes and language from the movie The Terminator to capture common fears of machines running amok and endangering humans, and possibly the human race.

Hollywood depictions aside, it is important not to lose sight of DOD's strong commitment to safety as a standard practice, including setting rigorous requirements, performing test and evaluation processes, and conducting legal reviews to ensure that weapons comply with international law. Consistent with this stance, safety holds a primary place in the DOD AI strategy, with one of the four lines of effort being AI ethics and safety. That said, in the current environment within DOD and in public discourse, safety can appear to be the forgotten stepchild of ethics. Ethics as a term is used to cover the broad set of concerns regarding military use of AI. The DOD's Defense Innovation Board is developing a set of ethical principles to guide its use of AI, and the Joint AI Center has discussed the need for ethicists on its staff—both are important steps for responsible use of AI. The military's use of AI must also be consistent with American values and principles.

However, what can be lost in the focus on ethics is that issues of AI safety—including operational outcomes such as civilian casualties, fratricide, and accidental escalation leading to international instability and conflict—are also significant when considering military use of AI. An ethics focus will not address these issues in a satisfactory manner; the appropriate set of experts for an ethics discussion is substantively different from the expertise needed for pursuing AI safety. This report shows that such a safety discussion will need to be more technical and operational in nature. Given this distinction, where is safety in DOD's initial steps? While recognizing both AI safety and ethics as key parts of its

strategy, DOD has not yet made the same kind of concrete institutional steps to reinforce the goal of AI safety.

The non-existent threat of Terminators aside, AI safety is a conversation that is vital for DOD to have. We note that DOD has strong policies and practices in place for safety when incorporating new technologies. At the same time, there are also strategic reasons to pursue AI safety specifically. These reasons include the following:

- AI is fundamentally different from other technologies in its abilities, requirements, and risks.
- DOD is operating in a different environment where cooperation with industry is vital and safety is a particular concern in that relationship.
- There is a need for appropriate trust in this new technology, which requires considering and managing safety risks and avoiding the extremes of overly trusting the technology and eschewing its use.
- To maximize the asymmetric advantage of alliances in an era of great power competition, it is important to align efforts and be interoperable with allies, including addressing their safety concerns.

In this report, in light of the Navy's stated commitment to using AI, and given the strategic importance of AI safety, we provide the Navy with a first step toward a more comprehensive approach to safety. We use a risk management approach to frame our treatment of AI safety risks: identifying risks, analyzing them, and then suggesting concrete actions for the Navy to begin addressing them. This includes two sets of safety risks, those associated with the technology of AI in general, and those associated with specific military applications of AI in the form of autonomy and decision aids. The first type of safety risk, being technical in nature, will require a collaborative effort with industry and academia to address effectively. The second type of risk, being associated with specific military missions, can be addressed with a combination of military experimentation, research, and concept development to find ways to promote effectiveness along with safety. For each type of risk, we use examples—bias for the first type and autonomy for the second type—to show concrete ways of managing and reducing the risk of AI applications. We then discuss institutional changes that would help promote safety in the Navy's AI efforts.

This action plan gives the Navy a starting point to effectively identify and manage safety risks in its applications of AI. We recommend a number of actions for the Navy, including the following:

- Given that AI is a technology poised to revolutionize warfare, Navy leadership should convene a board that looks at the potential applications, benefits, and risks of AI, challenges common assumptions about the future operating environment and threats, and develops recommendations regarding a set of potential operating concepts to explore and pursue.

- The Navy should conduct regular experiments to test and explore new operating concepts designed to leverage AI. These experiments should include a focus on exploring concepts for human-machine teaming and autonomous systems.
- These experiments should be coupled with the refinement of operating concepts to help the Navy adapt to a new way of warfare, providing feedback to help the Navy leverage the strengths of AI. These operating concepts should promote effectiveness and address safety issues discussed in this report.
- The Navy should develop a rigorous assessment process to enable effective learning from experiments, provide feedback to concept development, and inform capability requirements. These assessments should pay close attention to identifying ways to optimize the human-machine team and cultivate appropriate trust, including ways to manage safety risks introduced by the use of AI.
- The Navy should work collaboratively with academia and with industry on the identified general safety risks of AI: fairness and bias, unpredictability and unexplainability, and cyber security and tampering. This collaboration can strengthen the safety enterprise of the Navy, including setting system requirements for safety, improving test and evaluation processes, refining the conduct of legal reviews, and, for autonomous systems, informing senior level reviews as required for DOD Directive 3000.09.
- The Navy should designate a point of responsibility and take a deliberate a risk management approach for AI applications, including efforts to establish context, identify risks associated with AI technology in general and specific planned applications of AI, analyze and evaluate those risks to inform their scope and severity, and then determine ways to address them.

This page is intentionally blank

Contents

Avoiding SKYNET

In the Terminator series of movies, the Department of Defense (DOD) leverages Artificial Intelligence (AI) computer systems developed by Cyberdyne Systems to develop a learning system for automated command and control (C2) functions, called SKYNET. Cyberdyne developed its computer systems based on AI technology from the future that could think adaptively as humans can. When SKYNET becomes self-aware, DOD tries to shut it down, leading to SKYNET reacting in self-defense: launching a nuclear strike at Russia and drawing retaliatory strikes that destroy those trying to turn off SKYNET. This event, on August 29, 1997, later referred to as Judgment Day, results in the death of 3 billion people. After that initial cataclysmic event, SKYNET constructs battle robots called Terminators to eliminate the remaining humans, especially the emerging human resistance. This battle between men and machines moves into the past as SKYNET develops time travel and decides the best way to destroy the Resistance is to eliminate its leadership at an earlier time in history.[1]

The specific details, including the date of Judgment Day, change somewhat from movie to movie because of the paradoxes of time travel (and illustrating the challenge of developing a cohesive narrative in a series of movies with different directors). However, the overall narrative of a war between man and machine has proved to be compelling in American culture, especially as speculative science fiction catches up with current science fact. In 1985, for the original movie, AI was an abstraction only achievable in the future, with examples of the technology only evident because of time travel. Today, AI is a technology used increasingly in many areas of our lives, including navigation, banking, medicine, and integrated functions in our smart phones and household devices. This technology has proved to be so powerful that private companies are devoting tremendous Research and Development (R&D) resources to develop new applications. As predicted in the movies, AI is indeed a powerful and versatile technology.

The series also predicted that the technology would be of interest to militaries, which has also turned out to be true. With new, powerful AI applications in other areas of life, it is not surprising that nations have noticed and are planning to take advantage of the technology for their own interests. More and more countries around the world see advances in AI as central to their security strategy. For example, Russian President Vladimir Putin has remarked that

[1] SKYNET, Terminator Wiki, https://terminator.fandom.com/wiki/Skynet.

"the one who becomes the leader in this sphere [AI] will be the ruler of the world."[2] Similarly, China's plan to become world leader in AI has been compared to a 21st-century equivalent of the national commitment of resources embodied in the US mission to the moon in the 1960s.[3] In 2017, China's central government released the Next Generation Artificial Intelligence Development Plan, which explicitly seeks to "promote all kinds of AI technology to become quickly embedded in the field of national defense innovation."[4]

The US has also stated that leveraging AI is key to its new national military approach: the "Third Offset Strategy." According to this strategy, the US effectively leveraging AI technology is critical to keeping a military edge over potential adversaries and providing an effective deterrent to major conflict. The recent Executive Order (EO) on AI and DOD AI strategy also underscore the strategic importance of AI to the US government and military. The EO defines primary goals for US government use of AI and establishes primary responsibilities for key actions, and the DOD AI strategy establishes the new Joint AI Center and lays out four lines of effort to leverage AI more successfully.[5]

Hollywood has affected how these efforts are perceived. The government and military emphasis on the importance of AI has raised concerns by many about what this will look like in practice. Images such as the Terminator have dominated media coverage of the topic, and the development of Terminators and SKYNET are often mentioned in discussions about the ethics and safety of AI in military applications. DOD leadership has called the role of AI in autonomous weapons the "Terminator conundrum." Recently, in *Defense News,* a Navy official was quoted regarding Navy use of autonomous weapons and AI, with a headline of "The US Navy says it's doing its best to avoid a 'Terminator' scenario in quest for autonomous weapons."[6] Predictably, the public did not respond favorably to the Navy's mild assurance that it is "doing its best" not to end the human race. One such response on Twitter is shown below.

[2] James Vincent, Putin says the nation that leads in AI 'will be the ruler of the world,' The Verge, September 4 2017, https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

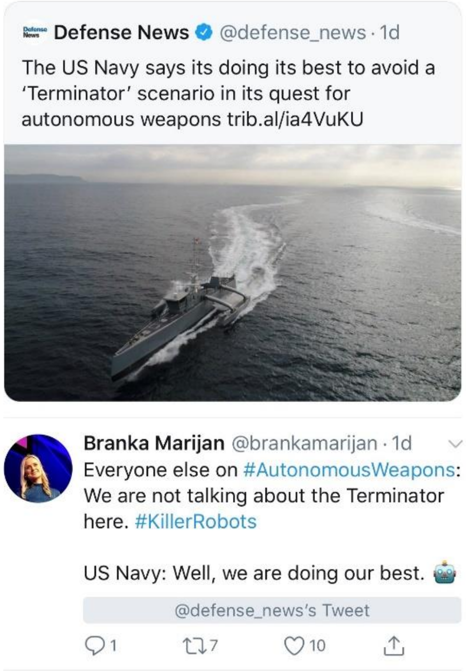
[3] Colin Clark, Our Artificial Intelligence 'Sputnik Moment' is Now: Eric Schmidt & Bob Work, Breaking Defense, November 1, 2017, https://breakingdefense.com/2017/11/our-artificialintelligence-sputnik-moment-is-now-eric-schmidt-bob-work/.

[4] Gregory C. Allen, Understanding China's AI Strategy, CNAS, February 6 2019. https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

[5] The White House, 'Executive Order on Maintaining American Leadership in Artificial Intelligence,' 11 February 2019, <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/; Department of Defense, Summary of the 2018 Department of Defense Artificial Intelligence Strategy, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>.

[6] https://www.flightglobal.com/news/articles/dod-official-world-faces-terminator-conundrum-on-421094/.

Figure 1. Perception of Navy AI efforts affected by Hollywood



Defense News @defense_news · 1d

The US Navy says its doing its best to avoid a 'Terminator' scenario in its quest for autonomous weapons trib.al/ia4VuKU

Branka Marijan @brankamarijan · 1d

Everyone else on #AutonomousWeapons: We are not talking about the Terminator here. #KillerRobots

US Navy: Well, we are doing our best. 🤖

@defense_news's Tweet

Source: Twitter.

Despite these Terminator references, in reality many of the fears about military use of AI are based on inaccurate perceptions of the state of the technology. DOD could not build a SKYNET system even if it wished to—and DOD leadership and policy have repeatedly stated that such a development is not consistent with its interests and values. The world can rest easier knowing that the Navy is not going to end the world in its pursuit of advanced technologies as it seeks a military edge.

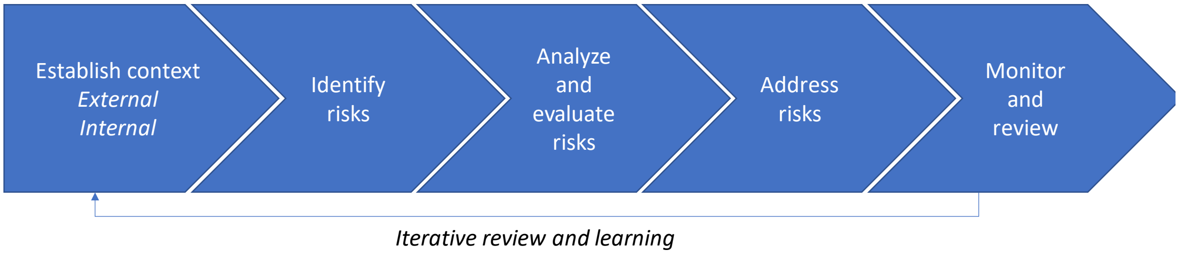
The non-existent threat of Terminators aside, it is vital that the Navy have a conversation on AI safety. Although the Navy has strong policies and practices in place for safety when incorporating new technologies, there are strategic reasons to pursue AI safety specifically. In this report, we provide the Navy with an action plan for pursuing AI safety.

We have used a risk management approach to frame our treatment of this issue. (Figure 2 illustrates a common risk management approach.[7]) This report aims to contribute to the Navy taking a rigorous approach to managing safety risks of AI by addressing key components of the

[7] International Standards Organization, Risk Management, ISO 51000, 2018, https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100426.pdf.

risk management framework, thus providing a head start for larger Navy efforts that are required to identify and manage applicable safety risks comprehensively.

Figure 2. Framework for risk management



Source: Risk management steps taken from ISO 51000.

We begin by establishing context, starting with a discussion of strategic imperatives of AI safety that increase the importance of addressing safety risks. We then continue considering context by discussing the technology and specific types of AI applications that the Navy is most likely to leverage, which helps us to frame the kinds of risks that AI can present. We then identify specific risks for each type of application, analyzing them and highlighting ways to mitigate them. We conclude with concrete steps that the Navy can take to effectively manage the risks that AI can create.

We note that such an approach is consistent with Stephen Hawking's warning of the exigent risks that AI presents to humans: "All of us should ask ourselves what we can do now to improve the chances of reaping the benefits and avoiding the risks."[8]

[8] Stephen Hawking, Stuart Russell, Max Tegmark, and Frank Wilczek, Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?', The Independent, May 1 2014, https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

Strategic Imperatives for AI Safety

The US Navy, and DOD as a whole, treat safety in military capabilities seriously and deliberately. In conversations between DOD and Silicon Valley, this point has surprised tech experts, who are accustomed to the approach taken in industry, where rapid fielding and rush to failure are integral to commercial success. In contrast, DOD processes are slow and deliberate, including Test and Evaluation (T&E) processes and legal reviews, because the business of DOD includes the delivery of lethal force and the handling taxpayer dollars in a responsible manner.

The consideration of safety with the integration of new technology to operations is not unique to AI. Weapon system use in warfare is governed by law—specifically the law of armed conflict as spelled out in the Geneva Convention. This body of law requires that militaries use a weapon or system in a way that meets the legal requirement to exercise the principles of military necessity, discrimination, and proportionality. The US has processes and policies in place to ensure that this is the case for all military capabilities, including AI.

Given these considerations, are there special attributes of AI that warrant additional attention? Part of the public concern about AI safety for military applications stems from a misunderstanding about the state of AI. For example, fears such as those expressed by Elon Musk are based on general AI, which is not present today or likely to be in the near future, if ever.[9] However, other concerns about AI applications are reasonable given the state of the technology. The Navy will naturally pay attention to AI safety because it is committed to legal and ethical conduct, including in the use of force, and the use of AI is no different in that regard.

However, AI does present several special considerations that warrant the Navy emphasizing AI safety. In the framework for risk management, these considerations can be considered as part of the first step of managing risk—establishing context. That step "defines the external and internal parameters to be taken into account when managing risk," with two elements,

[9] General AI is a hypothetical type of AI that can perform broad sets of tasks, understand, and learn as humans do. For example, Elon Musk's comments on the risks of general AI include: "If one company or small group of people manages to develop god-like superintelligence, they could take over the world." He continued: "At least when there's an evil dictator, that human is going to die. But for an AI, there will be no death—it would live forever. And then you would have an immortal dictator from which we could never escape." Cadie Thompson, "Elon Musk Just Issued a Nightmarish Warning About What Will Really Happen if AI Takes Over," Science Alert, April 6, 2018, https://www.sciencealert.com/elon-musk-warns-that-creation-of-god-like-ai-could-doom-us-all-to-an-eternity-of-robot-dictatorship. This is not to say that such risks should be downplayed. Rather, this should be a separate discussion on future threats compared to immediate and possible near-term applications.

external context and internal context.[10] Here we discuss external context, namely, external factors relevant to meeting mission objectives. These factors help determine the potential impact of risk factors, thus informing the prioritization of efforts to help evaluate and address risks. For application of AI to the Navy, these considerations include the following:

- AI is new to the military
- The military has a new dependence on industry
- The military needs a "Goldilocks" trust in AI
- The needs to consider working with allies in its use of AI

We discuss each of these considerations in turn.

AI is new to the military

History is replete with examples of militaries leveraging technology for military advantage. A few examples include:

- The chariot. The first vehicle on the battlefield, the chariot was adapted for speed and mobility from a common-use cart, resulting in a significant military advantage. Chariots have been described as "the superweapons of their age."
- Gunpowder. An accidental discovery, gunpowder revolutionized the waging of war by allowing militaries to harness the explosive force of a chemical reaction for speed and power.
- The crossbow. The invention of the crossbow, with its amplified power and improved accuracy over the bow and arrow, equalized the battlefield in an era of armored aristocracy. This leveling effect was of such concern to established powers that the church banned its use in some contexts.
- The internal combustion engine. This engine extended the advantages of the steam engine, changing the speed and reach of war. A few applications include powering logistics (supply trucks) and persistent and long-range surveillance and strikes from submarines, aircraft, and missiles.

Most introductions of technology change warfare for a time. There are a few that have radically changed the shape and scope of warfare, including gunpowder and nuclear weapons. AI is considered to be in that category. It brings applications across the enterprise of war, allowing both greater effectiveness and greater efficiencies.

[10] Chartered Accountants, Establish the Context: Risk Management, https://survey.charteredaccountantsanz.com/risk_management/small-firms/context.aspx.

AI is also different because of its unique characteristics. We first note that real-world applications of AI will be applications of narrow AI versus general AI for solving specific problems. We will discuss the two kinds of AI—with only narrow AI existing today and in the near future—in the next chapter, so for the moment we make the caveat that for AI that actually exists, some characteristics of AI are fundamentally different from existing technologies. For example, machine-learning techniques using large data sets essentially find optimal curve fits within a complex, multidimensional space in ways that we cannot envision or understand. This has implications, for example, for trust and certification processes that will require a new approach to military processes such as T&E.

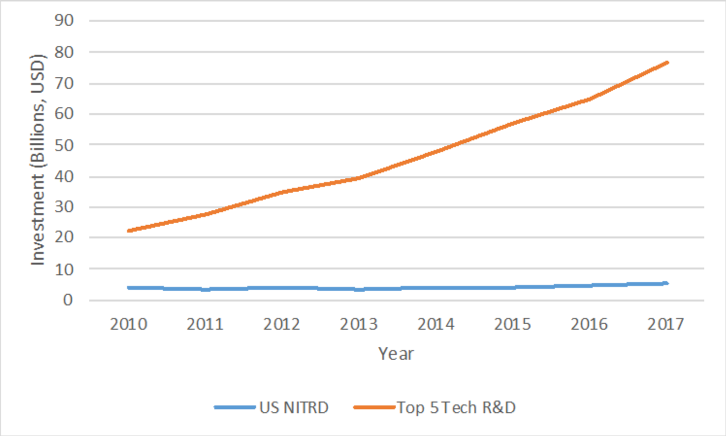
The military application of AI is analogous to the historical example of the US military's incorporation of nuclear weapons in this way: the critical technical knowledge necessary for safety was held largely by the civilian sector. Similarly, the technical knowledge regarding AI that the Navy will need to understand and respond to is possessed largely by the tech sector.

A new environment with industry

The dramatic advances in AI also create a new dependence on industry for the Navy. Since the Second World War, the US government has depended heavily on its own R&D investments. However, R&D investments in AI are increasingly dominated by the private sector, marked by a dramatic increase in R&D spending by the tech industry over the past decade. In Figure 3, we contrast the Networking and Information Technology Research and Development spending for the entire US government with the R&D investments by the top five US tech companies (Amazon, Google/Alphabet, Intel, Microsoft, and Apple). The tech sector invests much more in R&D, and the gap is increasing, as shown in Figure 3. The tech sector's R&D spending in 2010 represented six times the overall technology R&D investment of the US government. Eight years later, the tech investment spending grew to15 times that of the US government. Overall, the US faces a rapidly growing gap in research investment in cutting-edge technology. This creates a changing environment for the US in which cooperation with industry is vital for maintaining the technological advantages needed to meet US strategic goals. To the extent that AI safety is a concern to industry—and it is, as evidenced by companies such as Google eschewing support to US military applications and starting ethics review processes to monitor their internal processes—then the Navy will need to work with industry to help address such concerns.[11]

[11] This paragraph is adapted from an upcoming article, Resolving the Battle over Artificial Intelligence in War, to be published in the RUSI Journal later this year.

Figure 3. Growing R&D investment gap between US government and the tech sector



Source: Larry Lewis, Resolving the Battle over Artificial Intelligence in War, RUSI Journal, pre-publication draft, September 10 2019.

The military needs a "Goldilocks" trust in AI

AI safety also relates to trust. A key question regarding US military use of AI is whether military personnel and US government senior leaders trust that these systems will be effective and not cause inadvertent problems. The 2016 Defense Science Board study on autonomy made this point: "The individual making the decision to deploy a system on a given mission must trust the system."[12] Operations in Iraq and Afghanistan illustrate that commanders and operators responsible for a given operation will not necessarily use a system when they do not fully understand what the effects will be. When systems were fielded to meet urgent needs—such as counter-improvised explosive device systems or surveillance systems providing critical intelligence—some forces chose weapon systems and intelligence, surveillance, and reconnaissance platforms that they were already familiar with, even when they were inferior to newer capabilities available to them.[13]

Having too little trust in AI systems is one danger, as it prevent forces from using capabilities they need. Another danger is having too much trust in a capability. Humans have a propensity to overly trust machines, even when they have evidence suggesting that such trust is not

[12] Defense Science Board, Report of the Defense Science Board Summer Study on Autonomy, Washington, DC: Office of the Secretary of Defense, June 2016, https://www.hsdl.org/?view&did=794641.

[13] Larry Lewis, Insights for the Third Offset: Addressing Challenges of Autonomy and Artificial Intelligence in Military Operations, September 2017.

warranted. For example, in an experiment conducted by Georgia Tech Research Institute, 40 test subjects were directed to follow a robot labelled "Emergency Guide Robot." At first, the robot served as a guide as the subjects went through an unfamiliar building. The robot exhibited erratic and unreliable behavior, such as going in the wrong direction, moving in circles, or simply shutting down. However, when the subjects were put in a seemingly emergency situation, they all followed the robot, trusting it to lead the way to safety.[14] The researchers were surprised how willing the test subjects were to follow an unreliable machine.

There are concrete cases of excessive trust exhibited in warfare. For example, in the Army PATRIOT shootdown of a Navy F/A-18 aircraft in Iraq in 2003, the system misidentified the aircraft as a tactical ballistic missile and made a recommendation to the operator to fire interceptor missiles in response. The operator approved the recommendation without independent scrutiny of the information available to him. Similarly, an airstrike in eastern Nangarhar Province, Afghanistan, resulted from a single source of technical-based information that had been reported previously and was not revisited or scrutinized prior to the strike, in which led to dozens of civilian casualties as US aircraft mistakenly targeted a wedding party.[15] This suggests that in operations, the military needs a "Goldilocks" trust in AI, not too hot and not too cold: both extremes of trust should be avoided. The goal is *appropriate* trust, with human involvement in decision-making that is based on experience and knowledge regarding capabilities and system performance.

Senior military and government leaders also influence the nature of military operations, including the use of specific technologies in war, through policy decisions. These policies can influence levels of oversight (for example, DOD Directive 3000.09 requiring a senior-level review of some types of autonomous systems), types of allowable technology used in warfare (e.g., restrictions on white phosphorus, a cluster munitions directive setting requirements on characteristics needed for such weapons to be used), and policy parameters in certain types of operations. For example, the 2013 Presidential Policy Guidance (PPG) and its 2017 replacement outline an approval and oversight process for some counterterrorism operations. These policies help ensure that military activities are consistent with US principles, values, and interests. These policy-level determinations tend to reflect the level of trust held for the reliability of these systems or types of operations. Note that all of these examples involve safety: DODD 3000.09 is designed to avoid "inadvertent engagements" (e.g., fratricide and

[14] Mary Beth Griggs, People Trust Robots to Lead Them Out Of Danger, Even When They Shouldn't, Popular Science, March 1 2016, https://www.popsci.com/people-trust-robots-to-lead-them-out-danger-even-when-they-shouldnt/.

[15] Abdul Waheed Wafa and John F. Burns, U.S. Airstrike Reported to Hit Afghan Wedding, New York Times, November 5 2019, https://www.nytimes.com/2008/11/06/world/asia/06afghan.html.

civilian casualties). Restrictions on white phosphorus and cluster munitions are also intended to reduce the risk to civilians from the use of these weapons, and the 2013 PPG included civilian casualties as a no-go criterion for the default approval of operations.[16] Thus, it can be expected that safety will be a part of future senior leader guidance and direction regarding the application of AI to warfare.

The needs to consider working with allies in its use of AI

Alliances and coalitions can be considered as asymmetric advantages, elements of an offset strategy against peer competitors such as China and Russia that are expected to operate unilaterally. Working within a coalition offers a number of benefits to US military operations. One is that a coalition provides a greater collective mass in terms of forces available compared to each nation's individual contribution. Another important benefit is greater legitimacy regarding an operation, including the right to wage war. Coalition partners can also bring complementary capabilities not possessed by US forces, aiding the conduct of operations. For example, in coalition operations in Helmand Province, Afghanistan, UK forces brought aviation platforms and intelligence capabilities that complemented US Marine Corps forces in targeting and air support.[17]

However, operating with allies also creates challenges, which we refer to as friction points, that are not present in a unilateral operation. These friction points can increase the costs and risks to coalition forces while reducing operational effectiveness. Some friction points involve institutional military forces—such as differences in force generation, interoperability, military culture, and information sharing across computer information systems—while others involve national policy such as differing ROEs, detainee policies, treaty commitments, and other national caveats.[18] This is a particularly critical area for maintaining freedom of action and interoperability as some allies consider pre-emptive legal bans or regulation of autonomous weapons and other military applications of AI. Collectively, these friction points complicate integration efforts and reduce unity of effort among coalition partners. The US will need to

[16] Larry Lewis, Insights for the Third Offset: Addressing Challenges of Autonomy and Artificial Intelligence in Military Operations, September 2017.

[17] Alexander Powell, Larry Lewis, Catherine Norman, and Jerry Meyerle, Summary Report: U.S.-UK Integration in Helmand, CNA, February 2016.

[18] Alexander Powell, Larry Lewis, Catherine Norman, and Jerry Meyerle, Summary Report: U.S.-UK Integration in Helmand, CNA, February 2016.

work proactively on addressing both institutional and policy friction points to capture the asymmetric advantages of allies and coalitions.

A related topic is export policies and processes. The US is the world's largest provider of military weapons and equipment, in accordance with both economic and strategic goals. Strategic objectives include maintaining alliances and common interests and maximizing interoperability in support of combined operations. With the development of military capabilities employing AI and autonomy, the US will need to decide what capabilities to provide to allies, and what systems to support with US components. Besides the issue of proliferation, there are policy considerations for safety and reputational risk: for example, an indigenously developed system for using lethal force against people, which has few or no safety measures, using a key US-produced component, could endanger civilians and imperil the reputation of the US if some mishap were to occur. A policy review process for such exports will be required to manage these risks.[19]

Conclusion

Safety is always important for the US Navy, and existing processes, laws, and policies help promote safety in the conduct of military operations. That said, the application of AI to warfare presents special challenges, meaning that the Navy will need to provide extra attention to promoting safety in its applications of AI. In the next section, we discuss internal context, examining the characteristics and specific types of applications of AI relevant to the Navy.

[19] We note that such a review process will also need to address the question of preventing critical technology from falling into the hands of adversaries. In the 1970s and 1980s, the US's Second Offset strategy had the denial of key technologies to potential adversaries as a primary element. Similar export restrictions are desirable with military applications of AI, a central part of the Third Offset, but in practice this will be more difficult to implement given the dominance of the commercial sector on technological developments.

Examining Navy Applications of AI

The previous section examined external factors influencing the impact of potential safety risks. This section addresses internal factors regarding risk, examining specific functions and applications that help determine the kind and severity of safety risks. This method is consistent with a basic principle of risk management: to manage risk, you must first identify specific sources of risk. For the Navy and its use of AI, we examine two considerations that introduce risk: fundamental characteristics and limitations of AI, as illustrated by current commercial uses and Navy priorities for applications of AI.

Commercial AI applications

AI has been defined as a machine with a "quality that enables an entity to function appropriately and with foresight in its environment." There are two classes of AI that are commonly discussed: "narrow AI" and "general AI." The dramatic achievements seen in the commercial world are all examples of narrow AI, which is used to execute specific functions in limited, well-defined contexts. One subcomponent of AI is machine learning, a set of techniques "designed to detect patterns in, and learn and make predictions from data." This approach allows software applications that can solve problems without explicit programming, by finding patterns in large data sets. We note that the effectiveness of machine learning relies on not only an effective algorithm design but also on the quality and robustness of its training data.

The past few years have seen powerful and wide-ranging commercial applications of AI. Some examples include:

- Transportation: AI powers navigation apps such as Google Maps and Waze as well as ridesharing software, including Uber and Lyft.
- Banking and fraud detection: Banks can identify potentially fraudulent patterns and raise alerts concerning questionable transactions. AI is also used to interpret handwriting in mobile check deposits.
- Making recommendations: Shopping sites (e.g., Amazon), social media sites (e.g., Facebook), and entertainment sites (e.g., Netflix) analyze user preferences and suggest other content based on observed patterns.
- Virtual personal assistants: Alexa, Siri, and other applications feature voice recognition and the ability to provide requested content in conversation with users.
- Improved medical diagnoses: AI has been seen to improve the accuracy and timeliness of diagnoses from medical scans.

These applications are powerful but their uses of narrow AI are also limited: they are only applicable to the specific problem for which they are designed. If the problem is changed even slightly, a narrow AI application cannot adapt. That adaptation is the province of "general AI," a technology that can operate in and adapt to undefined and dynamic scenarios and problems. Potentially, such a technology can far exceed human performance in a broad set of tasks. This is the kind of technology Elon Musk warns about—enabling an evil AI dictator—but it should be noted that this technology has been discussed since the early days of the development of AI technology and experts agree we are still decades away, if it is possible to achieve at all.

The powerful commercial applications of AI are varied, but they all represent a discrete number of functions applied to a wide variety of contexts. Common AI functions in the commercial world include the following:

- Automating tasks. AI can be used to automate routine functions, enable autonomous functions, and operate in man-machine teaming to reduce the time and burden on personnel for performing tasks (e.g., taking over administrative functions, improved medical diagnoses)
- Processing complex or large data sets. AI can enable analysis of larger data sets or additional data sources (e.g., voice and image recognition) to find solutions
- Predicting behavior. AI can reference features in past data to anticipate possible future behavior (e.g., Google Maps predicting traffic times)
- Flagging anomalies or events of interest. AI can be used to identify indicators of potential problems or events of interest and create alerts (e.g., banks predicting potentially fraudulent transactions)
- Data tagging and error correction. AI can recognize content and create tags so that data can be more effectively or efficiently exploited. This approach can also improve data quality (e.g., Facebook automated image tagging).[20]

These functions can be applied in different contexts and combined to obtain a wide variety of outcomes. The power and versatility of AI explains the rapid growth of commercial R&D in this technology.

Military AI applications

The power and versatility of AI has not gone unnoticed by the US government. Recognizing the value of AI to the US as a nation, the president issued the "Executive Order on Maintaining

[20] The descriptions of these AI functions are adapted from: Larry Lewis, AI and Autonomy in War: Understanding and Mitigating Risks, CNA, August 2018.

American Leadership in Artificial Intelligence" in February 2019. This order cites the promise and transformative nature of AI to justify the need for executive action. The EO recognizes that AI is a transformative technology with many different applications, aiming to use AI to "…drive growth of the United States economy, enhance our economic and national security, and improve our quality of life."[21]

The US military has a much more specific mission: US national security. As such, the applications of AI it seeks are appropriately mission-focused. DOD has characterized AI applications into two basic types: "at rest" and "in motion" capabilities.[22] "At rest" capabilities are systems that "operate virtually, in software, and include planning and expert advisory systems," while "in motion" capabilities "have a presence in the physical world and include robotics and autonomous vehicles." The Department of the Navy has expressed interest in three types of AI applications: autonomous functions (AI in motion) two types of "at rest" applications: decision aids and optimization functions.[23] We describe each below.

Autonomous functions

Autonomous functions are those that do not require human control. A commonly discussed autonomous function is a lethal autonomous weapon system, but autonomy can also extend to non-lethal functions. For example, autonomous functions can include navigation, logistics, or allocation decisions regarding sensors or weapons. Autonomous functions do not require AI—they can be programmed in a rule-based way, for example—but the flexibility and effectiveness of these functions can be enhanced by leveraging AI. They can substitute for humans in repetitive or dangerous tasks, and can save manpower requirements or provide rapid response time in tasks where time is of the essence. Alternatively, autonomous platforms can be used in concert with humans, such as an autonomous wingman or logistics vehicle.

Decision aids

Decision aids combine the strengths of machines with the strengths of humans: leveraging technology's ability to process large data sets and filtering or finding patterns to leverage the human ability to gain understanding and apply context. In complex settings and tasks, the performance of a human-machine team can exceed that of either individually. A practical real-world application of this principle is Centaur chess: a practice pursued by Grandmaster Garry

[21] White House, Executive Order 13859, Executive Order on Maintaining American Leadership in Artificial Intelligence, February 11, 2019.

[22] Defense Science Board, Summer Study on Autonomy, Department of Defense, June 2016.

[23] Larry Lewis and Diane Vavrichek, An AI Framework for the Department of the Navy, CNA, August 2019.

Kasparov after he was beaten in 1997 by the AI-driven IBM computer Deep Blue. Exploring how man and machine could work together, it was found that amateur-level chess players working cooperatively with modest personal computers could outplay both the highest end computer chess programs and human Grandmasters.[24] The same approach to man-machine teaming can be used in military applications.

Optimization functions

AI can also be used to perform optimization functions. For example, AI can leverage large quantities of data and identify patterns to contribute to improved and predictive maintenance, help identify candidates for recruitment, and improve the performance of existing systems (e.g., improved tracking in an air defense system). These functions do not make independent decisions about physical actions, and they do not interact with humans on a real time basis, but they make real contributions to making military systems and processes more effective and efficient.

These three Navy AI applications will be achieved in practice by using one or several of the common AI functions mentioned earlier. Those common AI functions help us to identify some general safety risk factors associated with the military use of AI. In addition, we use the three Navy AI applications to help identify specific safety risk factors associated with that application. We discuss these risk factors in the next section.

[24] Sydney J. Freedberg Jr., "Centaur Army: Bob Work, Robotics, and the Third Offset Strategy," Breaking Defense, November 9, 2015, https://breakingdefense.com/2015/11/centaur-army-bob-work-robotics-the-third-offset-strategy/.

Identifying Safety Risks

The next step in risk management after setting context is identifying risks. In this section, we examine significant risks that AI can present in Navy applications. There are a number of possible negative safety outcomes from the use of AI. Significant safety concerns include:

- Inadvertent or improper engagements (e.g., fratricide and civilian casualties) during the use of force
- Unintended behavior that leads to escalation and conflict
- Inadvertent activities that can lead to collisions: with other military entities and with commercial or civilian entities

How could AI contribute to these safety concerns? What are the risk factors associated with AI? We begin by examining general safety risks associated with the technology, which could potentially be evident in all three of the Navy AI applications. These will also represent the set of risks associated with the application of AI for optimization. We then look at the two other Navy applications, autonomy and decision aids, and identify specific risks associated with each that can lead to the safety concerns listed above.

General safety risks from AI

AI has proven to be a powerful and versatile technology. That said, it does have shortcomings and vulnerabilities. Major concerns regarding AI tend to focus on fairness and bias, unpredictability and unexplainability, and cyber security and tampering. We discuss each below.

Fairness and bias

A commonly voiced concern about military applications of AI is bias and fairness. For example, will racial factors lead to some groups being more likely to be targeted by lethal force? Could detention decisions be influenced by unfair biases? For personnel, could promotion decisions incorporate and perpetuate historical biases regarding gender or race?

Such concerns can be seen in another area where AI is already being used for decisions regarding security: law enforcement. It is an unfortunate fact that, in the US, groups of people are discriminated against by law enforcement, even though they are protected by law and legal precedents against such treatment. For example, after the fatal shooting by law enforcement of Michael Brown, an 18-year-old African-American man in Ferguson, Missouri, in 2014, the

Department of Justice (DOJ) conducted an investigation of the Ferguson Police Department to determine if racial bias was evident in its law enforcement activities. The investigation found evidence for pervasive discrimination within the department with stops, warrants, arrests, and sentencing all more severe and disproportionate for the African-American population (DOJ, 2015).[25] Despite significant efforts to address discriminatory practices in law enforcement, episodes such as the DOJ investigation of the Ferguson Police Department illustrate that such biases persist.

There are concerns that unfairness and bias, including racial bias, in law enforcement can carry over to the use of AI methods. Currently, a number of law enforcement agencies in the US are using AI-driven applications for predictive policing and risk assessments for decisions such as parole. The concern is that pervasive bias in the criminal justice system introduces bias into the data used for these automated tasks. The argument states that any predictions developed by AI-driven approaches using that data would then be affected by this bias.[26]

When considering bias, it is important to note that not all biases are bad. Some computational approaches purposely build in biases to improve performance. For example, a voice-to-text translator may decide to assume the language it is interpreting is universally English. If the tool is used where English is spoken, then that bias results in the computational approach providing a better experience for the user and a more faithful translation of what was said. The human mind also uses bias to make its processing more powerful. For example, human vision fine-tunes its performance through the use of biases in what the brain should expect to see. Optical illusions are often cases where images exploit how the brain processes visual information.[27] Instead of seeking to avoid biases, we should be aware of the kinds of biases we want to avoid. Thus the Navy should be consciously aware of the possible deleterious effects of biases on AI applications.

Unpredictability and unexplainability

When Google's AlphaGo defeated Lee Sedol in the complex game Go, the Google engineers had no idea what their neural network program would do from move to move. Unable to predict

[25] Former FBI Director James Comey described how this discrimination was a tragic aspect of law enforcement in the United States: "At many points in American history, law enforcement enforced the status quo, a status quo that was often brutally unfair to disfavored groups." James B. Comey, Hard Truths: Law Enforcement and Race, remarks delivered at Georgetown University, February 12 2015.

[26] According to this claim, the bias can exist even if attributes such as race are not used explicitly because those attributes can be correlated with other factors that on first glance seem neutral.

[27] Tom Griffith, Bias is Always Bad (answering What Scientific Idea is Ready for Retirement?), The Edge, 2014, https://www.edge.org/response-detail/25491.

its actions, they knew only that the program would make every move it made according to what maximized its chance of winning. This comparable to a military commander issuing a mission command type order—go take the hill—and then tactical forces figuring out the best way to do so. The result of AI computational methods is a powerful capability for which the specific results cannot be perfectly predicted or explained. A mathematician working at Facebook AI Research described the situation: "The best approximation to what we know is that we know almost nothing about how neural networks actually work."[28]

A lack of predictability can create significant risks. For example, what if a military application of AI gives a result that was not anticipated and as a result, an aircraft moves across a contested border? "Surprise me" is not in the preferred lexicon of a military that has entire doctrinal publications on planning, and where an accidental move can result in escalation to war or people being killed. Explainability is also a concern. Using law enforcement as an example, if an AI application is used to determine whether a prisoner gets parole based on his likelihood of recidivism, and the algorithmic approach outperforms the decisions of human judges, is it a concern if no one can know why that decision was reached? How does this same situation work in a military context? What is the standard for explainability of administrative and promotion decisions? Of targeting decisions?

Cyber security and tampering

As effective as AI applications may be, by their nature they will also be especially vulnerable to tampering and manipulation by adversaries. The case often discussed is an autonomous unmanned vehicle that another party takes control of—a scenario presaged by an incident in 2011 in which an unmanned vehicle, a RQ-170, was taken over and captured by Iran —but the possibilities for tampering are much broader.[29] Besides interference with autonomy, decision aids could be affected when they fail to notice threats or suggest targeting solutions that would end in fratricide or civilian casualties (e.g., a commercial airliner). Although these efforts could aim to hack the combat system or decision aid itself, that is not necessary when applications are built and optimized with large data sets. Instead, an adversary could tamper with training data to build in compromises or hidden features undetected. The area of computer vision, for instance, is rife with examples of carefully placed pixels or patches on an image changing what object's classification, capitalizing on the sensitivity of neural nets to small and sometimes imperceptible changes in an image.[30] Such vulnerabilities could be used to misclassify entities

[28] Kevin Hartnett, Foundations Built for a General Theory of Neural Networks, Quanta Magazine, January 31 2019.

[29] John Keller, Iran-U.S. RQ-170 incident has defense industry saying 'never again' to unmanned vehicle hacking, Military and Aerospace Electronics, May 2 2016.

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, Intriguing properties of neural networks, ARXIV, 2013, https://www.arxiv-vanity.com/papers/1312.6199/.

purposely with malicious intent. This weakness introduces more rigorous security requirements for AI applications across its entire life cycle, to include protection of the integrity of the data and curation process.

Safety risks from Navy-specific applications

In addition to general safety risks that are a feature of AI use in general, two Navy applications of AI introduce their own kinds of risks that will need to be considered: decision aids and autonomy.

Decision aids

DOD has stated that human-machine teaming is essential to future competitiveness and maintaining a military advantage over competitors.[31] The Navy expects human-machine teaming applications of AI to yield some of the most valuable contributions of that technology. However, this is not a trivial task. Optimizing the performance of the human-machine team requires effective integration of both technical system performance and human factors. History shows that this integration of man and technology is not guaranteed, with historical safety failures often marked by breakdowns in this teaming. Earlier in this report, we discussed the Army PATRIOT shootdown of a Navy F-18 and how the man-machine interface led to the shootdown because of inappropriate trust. Another such example is USS *Vincennes* and its shootdown of the Iranian Airbus commercial flight, which also featured a breakdown in man-machine interface, with the Navy crew misreading the altitude and the Identification Friend or Foe (IFF) information for the commercial flight, leading to the tragedy. The collision of the USS *Fitzgerald* with a commercial tanker ship is another example where available digital information, specifically the Automated Identification System, was not leveraged to promote safety.[32]

Compared to those historical examples, the information available to Navy forces today is much richer and broader in scope. The risk of cognitive overload is considerable. While AI in decision aids can help to filter down information to allow appropriate and effective focus in a sea of information, there are critical unresolved issues. For example, how should the decision aid aim to present information to the decision-maker? What assumptions regarding the decision-making process and the cognitive load are implicit in that design? What are the optimal ways

[31] Cheryl Pellerin, Work: Human-Machine Teaming Represents Defense Technology Future, DEFENSE.GOV, November 8 2015.

[32] Geoff Ziezulewicz, The Ghost in the Fitz's Machine: why a doomed warship's crew never saw the vessel that hit it, Navy Times, January 14, 2019.

for the man and the machine to interface, in terms of senses and communication approaches? How much context should the decision aid provide, and is there capability to explore the assumptions behind automated recommendations?

There is much at stake with these questions. For the human to trust too little in the decision aid, there is a risk of inaction or using unaided human judgment to make decisions. In addition, real-world operations show repeatedly that humans make mistakes in the heat of battle. On the other hand, for the human to overtrust the machine has its own sources of peril. Thus, trust needs to be carefully calibrated.

Part of that trust is understanding that an AI-driven system is not actually thinking as a human does. These systems make decisions that humans cannot understand and explain. There is an internal logic based on the data provided, but the machine input is based on data and lacks context regarding what is important to the mission and what is acceptable from a human values and ethics perspective.

At the same time, in an environment where man and machine learn from each other, there are also risks in a situation where an AI system is trained by observing humans. This is because humans can have biases and bad traits. It is conceivable that an AI-driven decision aid that is learning from interactions with humans could learn the wrong things, creating risk.

Autonomy

Both international and domestic concerns over military applications of AI tend to focus on lethal autonomous weapon systems: military systems that can make engagement decisions without a human needing to pull the trigger. Concerns over risks of autonomous weapon systems include changing the strategic international balance of power that maintains effective deterrence, making war more likely, and empowering authoritarian regimes. The most commonly expressed safety concerns regard limiting undesired weapon effects (civilian casualties and fratricide) and inadvertent escalation leading to war.

If these are the overall safety risks, what are the specific risks of autonomous systems that can lead to those unintended outcomes? We have analyzed both fratricide and civilian casualty incidents, characterizing the common mechanisms for both types of incidents. Based on that analysis, we discuss the most common risks that lead to those outcomes that could apply to autonomous systems.

Fratricide risks

Deconfliction of friendly forces. In fratricide incidents, there was usually someone in the force who was aware the engaged entity was friendly. However, that information was not communicated to the shooter.

IFF. IFF capabilities have not always been reliable, with failures leading to greater risk of fratricide. In addition, there have been cases where a mismatch existed between the IFF type and the sensor used to detect it (e.g., orange identification panel and an IR sensor).

Poor situational awareness. Often in fratricide incidents, the shooter was unaware of friendly locations (e.g., Forward Line of Troops) or of his own location in relation to friendly locations. This meant that the shooter did not attempt friendly deconfliction because he was unaware of friendly forces in the area.

Civilian casualty risks

Unaware of civilians in the target area. There are several scenarios where the shooter may be unaware of civilians in the area of weapon effects. These include civilians moving into the area at the last minute (which can also happen after an initial strike when first responders move in to provide medical care) and civilians unobserved in attacks on buildings or vehicles where surveillance cannot detect them.

Misidentification of civilians as a valid military target. This can happen through misinterpretation of activity (considered a threat based on observed hostile act or intent) or by misassociation with a valid military target (e.g., they cross paths and there is confusion regarding which is which).

Not exercising tactical patience to reduce risks to civilians by choosing the optimal time and place for the use of force.

Not recognizing protected status of a target or entities in proximity to the target. This may include a medical facility, surrendering military forces, and vehicles on a humanitarian mission.

Inadvertent escalation

Another concern regarding AI safety and autonomous functions is the possibility of inadvertent escalation—where an autonomous platform commits an action (the use of force, movement across a border, or other potentially threatening behavior) that is then interpreted by an adversary as justification to escalate toward the use of force. Such functions can include autonomous navigation, the risk of accidental collisions, and the use of lethal or potentially less-than-lethal force.

Summary

In this section, we identified key safety risks regarding the use of AI, including general risks associated with the technology and more specific risks associated with Navy-specific applications. In the next section, we discuss how the Navy can address these risks, both specific risks discussed here as well as building a safety process to more comprehensively address safety risks associated with AI.

Addressing Safety Risks of AI: An Action Plan for the Navy

The next steps in the risk management framework discussed in the first section are to analyze, evaluate, and address risks. These tasks ask a number of questions, such as: What are the likely consequences of existing risk factors? What are the current processes in place to address those risks? Given that it is impossible to eliminate risks completely, what strategy can be used to bring risks to an acceptable level?

In the previous section, we identified several kinds of risks that AI can present within the context of military operations. These included safety risks—inadvertent engagements, accidental escalation to conflict, and accidents—and other risks such as bias and lack of explainability that will have broad impacts on the Navy as a whole if not addressed. Here we focus on mitigating safety risks, noting that doing so will also help address some of these other risks. We start with specific steps that the Navy can take to address general risks of AI identified in the last section, fairness and bias, unpredictability and unexplainability, and cyber security and tampering. We then discuss an approach to addressing risks in the two Navy AI applications that have their unique risks: decision aids and autonomy. Finally, we discuss institutional changes that the Navy should make more broadly to help it leverage AI more swiftly, effectively, and safely.

Addressing general risks of AI

The risks discussed in the previous section, fairness and bias, unpredictability and unexplainability, and cyber security and tampering, are symptoms of the nature of the technology. Some of them, such as the success of adversarial spoofing of image recognition from very small modifications of the image, were surprising even to the best experts in the field, showing that there is still very limited understanding of how this technology actually works in practice. These computational approaches often work powerfully, but we cannot explain why. Thus exploring these inherent risks of AI, such as the image recognition weakness, is a very technical venture. The Navy does not have the technical expertise necessary to conduct basic research to identify and clarify such risks. Thus, the current situation with the Navy and AI is comparable to the beginning of the US military's investigations into nuclear weapons. This is not to say that AI is as destructive as nuclear weapons, rather that deep expertise in AI is in the civilian sector, as it was with the advent of nuclear weapons.

The good news is that the private sector has much invested in AI performing well. As the top five tech companies are spending 15 times the total IT investment of the US government on R&D, part of this investment is dedicated to understanding the risks and weaknesses of AI. This is something that the Navy can pay attention to and benefit from. Similarly, academia is giving considerable thought to these questions, and the Navy can partner with academics to help address both general concerns about AI risks and specific applications that create safety concerns. We propose that the Navy look to insights and solutions from industry and seek to leverage academic work on the three risks we have discussed here: fairness and bias, unpredictability and unexplainability, and cyber security and tampering. To illustrate how the Navy could do this in practice, we provide some thoughts on how this process would work. To do so, we focus on fairness and bias, which was the greatest international concern regarding AI voiced during the 2019 UN talks on lethal autonomous weapon systems.

Addressing fairness and bias: using law enforcement as an example

In the previous section, we discussed how other sectors are ahead of the Navy in terms of applying AI in practical ways, and the example we highlighted was law enforcement. With lives on the line—both those who could have their freedom taken in incarceration and those who could be victims of crime—the stakes are high to make sound decisions, and cities around the US are turning to AI-driven methods. However, this kind of approach has also raised concerns about bias. For example, are certain groups discriminated against in this approach? How can we know if these methods are fair?

Illustrating how academia can help the Navy, academics have been considering the problem of fairness within the context of AI applications in law enforcement. When concerns of fairness arise, the first question should be: What do you mean by fair? For example, Richard Berk at the University of Pennsylvania has described a number of different kinds of fairness within the context of AI approaches that can be evaluated analytically.[33] They include the following:

- Outcome unfairness. Here, protected groups (or regional areas) are over-represented as high risk in the computational approach (e.g., a higher proportion of one group is seen to fail on parole). Assessing the performance of the computational approach, are some protected groups forecasted to have a greater probability of some consequential outcome than other protected groups?
- Classification unfairness. In this kind of unfairness, protected groups or specific areas are more likely to be misclassified either as high risk when they are not (i.e., a false

[33] Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research, 1-42, 2018.

positive) or as low risk when they are not (i.e., a false negative). Given the observed outcome for the computational approach, what are the chances of getting it wrong, and are there differences in these chances for protected groups?

- Forecasting unfairness. In another kind of unfairness, protected groups or areas are more likely to be subject to inaccurate forecasts. That is, they are more often incorrectly forecasted to "fail" or incorrectly forecasted not to "fail." Given the computational forecast, what are the chances of getting it wrong and are these chances different for protected groups?
- Treatment unfairness. In this kind of unfairness, protected groups or areas are treated differently by the model or algorithm. For example, in the computational approach, false positives are given less weight for one group compared with another. Given the same risk factors, what are the chances that they will be handled differently for different groups? Such differences affect the forecast made.

This shows that fairness is more complex than is commonly understood. A challenge in trying to make computational approaches "fair" is that, with these different dimensions of fairness, there are necessary tradeoffs that need to be made when seeking a particular type of fairness. Specifically, fine-tuning the computational approach to achieve one type of fairness will by necessity reduce fairness of the other types. At the same time, constraining the computational approach to ensure fairness of a specific type will also reduce its predictive performance. This is not to say that fairness is not a worthy goal. However, it should be understood that seeking fairness will involve trade-offs that should be explicitly addressed, with a common and precise understanding of what is meant by fairness in the context of the application.

That said, law enforcement computational methods also illustrate that these can serve as tools to better understand and address existing bias. For example, in the previous section we acknowledge that law enforcement in the US has a long history of racial bias. While some have concerns that AI-driven approaches to law enforcement will perpetuate those biases, these tools also have the opportunity to illuminate existing biases and help address them. For example, pre-processing of the input data could identify existing biases in law enforcement processes and decisions and explore ways to reduce them. This could include identifying problematic practices (e.g., stop and frisk) as well as officers and judges who seem to make decisions or arrests that may be compromised by bias. Modifying the input data accordingly, we can then explore whether there is a positive effect on the intended fairness measures. There are also ways to adjust computational approaches to constrain the algorithmic approach and ensure fairness: for example, using the statistics for one group believed to be free from bias and treating all cases with those statistics. Then the outcomes of these approaches can be evaluated analytically to determine how they comport with the aimed-for fairness measures. In such a way, AI—so often believed to be hopelessly bound to bias—can in fact be a tool to identify and correct existing biases.

This case illustrates an approach that the Navy can take to address intrinsic risks associated with AI, working closely with academia and industry to better characterize and then mitigate those risks. We note that DOD has announced initiatives on the other two risks we discuss here: unpredictability and unexplainability (addressed in part by a Defense Advanced Research Projects Agency project on explainable AI), and cyber security and tampering (which is one of the National Mission Initiatives of DOD's Joint AI Center). While the Navy should participate in those initiatives and learn from them, in all three of these areas the Navy (in concert with the other Services when possible) should also seek to partner with academia and industry to characterize and mitigate these risks.

Addressing Navy-specific risks of AI

In addition to the general safety risks of AI, two intended Navy applications of AI carry their own special risks, as described in the previous section. Here we discuss some ways that the Navy can help to mitigate those risks.

Safety in decision aids: strengthening human-machine teaming

The marriage of man and machine can be a powerful combination—as seen in the earlier example of Centaur chess, where amateur chess players combined with simple programs could defeat Grandmasters or super-computers—but this marriage is a difficult one in practice. Past tragedies on the battlefield, both fratricide and civilian casualties, often show how difficult it is for humans to work with machines to make the best decisions. In the middle of conflict that is anticipated to be both swiftly paced and dense with critical information, how can the human-machine be tailored to obtain operational Grandmasters?

For Centaur chess, this was achieved through experimentation. Starting with the idea that men and machine are better together, the next step was to try various implementations of that idea, and then fine tune the concept based on lessons and failures. If the Navy views human-machine teaming as a critical component of a successful future force, then it should be doing the same: looking for opportunities to experiment and learn. This experimentation will then need to feed into concept development and training processes.

What should these experiments look like? They can be excursions into different mission sets, to maintain a diversity of use cases and avoid optimizing human-machine approaches for a specific case or scenario. They can also exercise anticipated characteristics of future conflict, using a combination of live exercise or operational data and simulation to explore and stress the success of the interface and find areas of failure to fuel faster learning. They should also incorporate rigorous assessments, looking at performance in detailed, data-based evaluations instead of relying on observers to give a thumb's up or down. As the human-machine interface is extremely complex, such assessments should be designed to capture those complexities. Experiments can also explore different operating concepts, such as schemes for C2, operators controlling or monitoring a single platform, a collection of platforms, and self-organizing swarms, to understand the requirements and the risks associated with each.

These efforts should also be informed by cutting-edge research into human factors and technology integration. For example, based on research and recent developments, what should future experiments include with regard to how to present and filter information from machine to man? What are appropriate operator workloads in different settings? What technical and operational requirements are needed to be able to calibrate and develop appropriate trust in these systems? What risks and tradeoffs exist regarding mission effectiveness and safety, and how should those tradeoffs be managed?

The conduct of such experimentation and the learning and institutional changes that need to occur in response to them have larger implications for changes the Navy will need to make. We address these larger requirements for the Navy at the end of this section.

Safety in autonomy

Autonomous functions can carry risks, and this is particularly true for autonomous functions that can employ lethal force. A commitment to safety is why DOD created its Directive on Autonomy (3000.09) before any such systems were developed, aiming to be proactive and reduce the risk of inadvertent engagements (fratricide and civilian casualties). Recognizing this imperative for safety with autonomous weapon systems, how can the Navy pursue this in practice?

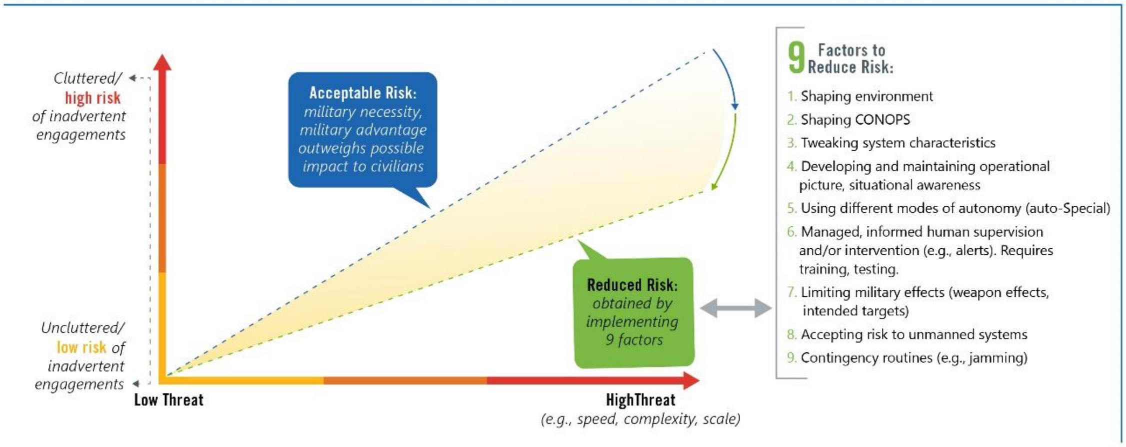
There are two components to safety of autonomous weapons. The first is addressing the technical is the risks, which we have described above: bias, predictability, and cyber security. The Navy addressing such risks will improve the safety of all of its uses of AI, including the function of autonomy.

The second component of safety is operational and organizational steps the Navy can take to promote safety in its use of autonomy. Figure 4 illustrates a number of such steps that the Navy can take to promote safety in autonomy. Autonomous weapons, like any weapon, must operate within the constraints of the laws of war (otherwise known as international humanitarian law). One element of the laws of war is the requirement of military necessity for the use of force, including that the military advantage must outweigh the potential impact to civilians.[34] This is represented by the upper blue line in the below figure, where a higher risk of unintended engagements may be acceptable in response to a greater military necessity for action based on

[34] Article 57, Protocol Additional to the Geneva Convention of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I), of 8 June 1977.

a higher threat. This shows that international law accepts a certain level of risk in the use of military force within the context of warfare.

Figure 4. Factors to reduce risk in the use of autonomy


Cluttered/
high risk
of inadvertent
engagements
Acceptable Risk:
military necessity, military advantage outweighs possible impact to civilians
Reduced Risk:
obtained by implementing 9 factors
Uncluttered/
low risk of
inadvertent
engagements
Low Threat
HighThreat
(e.g., speed, complexity, scale)
9 Factors to Reduce Risk:
1. Shaping environment
2. Shaping CONOPS
3. Tweaking system characteristics
4. Developing and maintaining operational picture, situational awareness
5. Using different modes of autonomy (auto-Special)
6. Managed, informed human supervision and/or intervention (e.g., alerts). Requires training, testing.
7. Limiting military effects (weapon effects, intended targets)
8. Accepting risk to unmanned systems
9. Contingency routines (e.g., jamming)

Source: CNA.

That said, for the use of force with autonomous systems, there are a number of ways to potentially reduce that risk; we discuss nine of them below:

- **Shape the environment.** There is a direct connection between the operational environment and risk of inadvertent engagements. One example encountered in Iraq and Afghanistan featured checkpoint operations, where US forces mistakenly engaged civilians in the belief that they were insurgent threats. One risk factor was the available reaction time. Careful consideration of the risk of such inadvertent engagements influenced the placement of US force positions and other measures to shape the operating environment, leading to more time to react and make better decisions.
- **Shape the concept of operations.** There are times that mission objectives can be met in various ways, and they may not all require the delivery of lethal force, or they may allow warnings to be issued prior to the use of force. One example is the potential use of autonomous systems to defend naval ports. Changing the stationing of sensors and weapon platforms and building in steps to warn and potentially de-escalate the situation can still meet the defense requirements while improving safety.
- **Tweaking the system characteristics.** Systems are built according to military requirements. However, those requirements do not always incorporate possible safety considerations. Using the example of port defense, building in the capability to

deliver warnings or to attempt to disable ships before using lethal force can improve the safety of such systems.

- **Developing and maintaining situational awareness.** A critical component of safety is maintaining situational awareness. Historically, inadvertent engagements feature a loss of situational awareness, leading to mistakes resulting in fratricide or civilian casualties. The moment of a potential engagement decision is the wrong time to try to develop situational awareness. Building a process that maintains situational awareness of who is who early on reduces the risk of inadvertent engagements.
- **Using different modes of autonomy as appropriate.** Current systems can have various modes with different levels of autonomy. For example, Aegis ships have an Auto Special mode that can conduct engagements without human intervention, while other modes can recommend engagements but require a human operator to approve them prior to weapons release. Tailoring the mode of autonomy depending on the environment and threat is a way to manage risk of autonomous weapons.
- **Managed, informed human supervision and intervention.** Having human supervision before and/or during the operation of autonomous systems can also reduce risk. Besides the approaches of operator in the loop (approving engagements before they occur) and operator on the loop (no approval needed but an operator can intervene and stop the engagement), other aspects of human supervision are also valuable. For example, human monitoring of the location of an autonomous system related to the known locations of friendly forces, civilians, and the enemy order of battle can aid in decision-making. If the environment changes in a way that affects the risk of inadvertent engagements, humans can intervene accordingly.
- **Limiting military effects.** Another way to reduce risk is to limit military effects through the use of weapons with reduced areas of effects, different weaponeering options, or limiting autonomous engagements to specific environments (e.g., only on known military routes, or avoiding urban areas) or target types (e.g. a specific visual or infrared signature of military vehicle).
- **Accepting risk to unmanned systems.** One contribution to civilian casualties in military operations is a reactionary engagement taken in self-defense that turns out to involve civilians. Autonomous systems can avoid this risk to civilians by accepting more risk, using tactical patience, or allowing the platform to move in closer to get a more accurate determination of whether a threat actually exists.
- **Contingency routines.** One discussed risk of autonomous systems is that adversary jamming can sever communications and force a scenario where such systems then make engagement decisions without the possibility of human supervision or intervention. This risk can be addressed by having the option of contingency modes. For example, a commander can decide what the risk is of jamming and follow-on risks

of autonomous engagements. If the risk is unacceptable, then the system can be programmed to return to a pre-programmed location or to conduct a secondary mission.

The overall result of applying these nine factors in the design and use of autonomous weapons is a reduction of risk of inadvertent engagements. These are factors the Navy can include in platform development, in the development of concept of operations employing autonomy, and in training to promote the safety of autonomous weapons.

Needed: a Navy enterprise approach to AI safety

This section has discussed specific steps the Navy can take to promote safety in its use of AI. Some of these steps should be taken in cooperation with industry and academia, while others are clearly the responsibility and purview of the Navy, with possible cooperation from the other services and with the Office of the Secretary of Defense. One of the critical questions that arises within a risk management approach is: Where do responsibilities lie to make needed actions to manage and mitigate risk? More broadly, what steps does the Navy need to take as an institution to address AI security? This depends on how important the issue of AI safety is to the Navy and DOD.

How important is AI safety?

Some may say: DOD has repeatedly said that it wants to move fast with regard to applications of AI, and safety is simply going to slow it down. Thus, safety should not be a priority. This ignores two factors. The first is the actual relationship between safety and effectiveness and the second is the acknowledged place of safety in DOD's AI strategy.

First, some believe that safety is in conflict with effectiveness; steps taken to promote safety will hold back the force and make it less effective. However, recent US operations have shown this is a misunderstanding of how safety can be pursued operationally. In Afghanistan operations, Special Operations Forces (SOF) responded to concerns regarding civilian casualties by re-examining its processes and challenges and looking for ways to improve its targeting. The result was a force that had a reduced rate of civilian casualties and a greater rate of mission effectiveness. Seeking safety did not hinder mission effectiveness; rather, those efforts found weaknesses in the targeting process that simultaneously reduced effectiveness and increased the risk to civilians. Addressing safety thus led to more operations achieving the mission. A commitment to safety can make the force better at what it does.

Second, safety holds a primary place in the DOD AI strategy. One of the four lines of effort for DOD's strategy on AI is ethics and safety. That said, in the current environment within DOD and in public discourse, safety can appear to be the forgotten stepchild of ethics. Ethics is currently serving as a general catchphrase that tends to cover the broad set of concerns regarding military use of AI. DOD has the Defense Innovation Board (DIB) developing a set of ethical principles to guide its use of AI, and the Joint AI Center is on record about wanting ethicists on its staff. These are important steps for responsible use of AI. However, where has safety gone in DOD's initial steps?

Analysis of public comments in the DIB events seeking to get input for the development of its AI ethics principles shows that safety is a key component of that conversation. While there were a broad set of concerns raised, the concerns of safety—civilian casualties, fratricide, and accidental escalation leading to international instability and conflict—were major components of that conversation. Ethics is a tremendously important consideration for the military's use of AI, making sure that any use of that technology is consistent with American values and principles. At the same time, the appropriate set of experts for that discussion is substantively different from that needed for discussing AI safety. This report shows that such a discussion will need to be more technical and operational in nature. While recognizing AI safety as a key part of its strategy, DOD has not yet made concrete institutional steps to reinforce that goal.

This is not altogether surprising. AI is a new warfare innovation, often compared to the invention of gunpowder and nuclear weapons in its tremendous likely impact to future warfare. Historically, militaries struggle with making the changes needed to pursue rapid and effective incorporation of innovation. This can lead to delays in effective implementation and operationalization of available technology, potentially losing a military edge to adversaries that were also seeking to capitalize on these developments. Historical case studies show that successful military innovation is not so much about the technology itself but how the military as an organization harnesses the technology. There are four historical best practices that can help guide the Navy in seeking effective and safe use of AI: leadership and responsibility, experimentation, concept development, and assessment.

Leadership and responsibility regarding AI safety

The primary determiner for successful innovation began with leadership focus. This included a determination that change was needed, followed by the championing of the needed intellectual development of the problems to be solved.[35] This often took the form of a board convened for the purpose, taking an initial idea, challenging common assumptions, and determining details of what needed to be done. Given that AI is a technology poised to

[35] Donn Starry, To Change an Army, Military Review, 63 No. 3, March 1983

revolutionalize warfare, such a board could examine the potential applications, benefits, and risks of AI, challenge common assumptions about the future operating environment and threats, and develop recommendations regarding a set of potential operating concepts to explore and pursue. Looking at precedents, one example of this kind of initiative was the Hogaboom Board, convened by the US Marine Corps in 1956 to re-examine service-level assumptions that led to a different way of organizing and equipping for the atomic age. If the impact of AI on the US military is comparable to that of nuclear weapons, another such board would help the Navy prepare for success in this new environment.[36]

Another attribute of successful leadership of innovation was that leadership was proactive in pushing needed institutional changes, not leaving needed changes to the status quo. This is because institutions tend to resist disruptive changes in their mission to continue to improve in incremental ways. Such leadership included but was not limited to senior leadership at the service level. Other leaders could and did have significant roles to play, especially in taking the initiative in executing the following steps—experimentation, concept development, and assessment—without waiting for a mandate.

For AI safety, it may be useful to designate a proponent specifically for safety, as the issues discussed here cross domains and span technical and operational considerations. Regardless of whether there is a single proponent for safety, the Navy should ensure that safety considerations are handled in a deliberate risk management approach, including identification, assessment, and addressing of potential risks created by AI applications.

Experimentation

Experimentation is another critical aspect to operationalizing technology rapidly and effectively.[37] One striking element of historical experimentation is that it did not depend on having the actual technology available. For example, the Germans and their development of the Blitzkrieg approach to the use of tanks used regular vehicles to simulate the intended movements and tactics that could be explored. That way, when Germany acquired the technology, they were prepared to use it rapidly. Similarly, when the Marine Corps sought to use helicopters for amphibious operations, they stood up a squadron and started developing tactics without a single helicopter available to them.[38]

It is important to note that these successful instances were not technology demonstrations, or even exercises where forces practiced to gain tactical proficiency. Rather, they were structured

[36] Terry C. Pierce, Warfighting and Disruptive Technologies: Disguising Innovation, Frank Cass, 2004.

[37] Mick Ryan, Human-Machine Teaming for Future Ground Forces, Center for Strategic and Budgetary Assessments, April 2018.

[38] Terry C. Pierce, Warfighting and Disruptive Technologies: Disguising Innovation, Frank Cass, 2004.

to test and refine operational concepts. As such, they were tightly coupled to two other activities we discuss here: concept development and assessment. The design of experiments exploring AI effectiveness should include considerations regarding AI safety. For example, an experiment exploring autonomous vessels conducting ship or port defense can include attention to the nine safety factors we discussed earlier in this section and determine the best mix of such factors to promote effectiveness and safety. This experimentation can then inform refinements of operational concepts and system requirements, e.g., a need for an additional source of information to enhance situational awareness or an additional sensor. Thus, experimentation can fuel an iterative process of improvement in effectiveness and safety regarding AI applications.

Concept development

In historical innovation, experimentation was closely coupled to concept development to allow rapid adaptation and refinement of operational concepts that would address emerging threats and effectively integrate technology. AI holds the promise of potentially disruptive new tactics on the battlefield, for example, widespread use of autonomous platforms that can operate under water undetected until needed, use of unmanned swarms, and use of adaptive communications and electromagnetic countermeasures that can evade conventional jamming techniques.[39] However, developing a system with a certain set of capabilities is not enough; the operational use of these and other applications rely on their integration into operational concepts.

Some of the most disruptive applications of AI—autonomy and human-machine teaming, including human augmentation—are the least clear in terms of what precise approach will be most beneficial. These will likely require more experimentation and refinement of operational concepts to fine tune operational approaches using that technology. Having a fast, tightly coupled loop with experimentation and assessment can help to accelerate learning and reach effective operational capabilities sooner. As with experimentation, these operational concepts should be developed in light of both effectiveness and safety.

Assessment

Assessment is a topic often neglected in current discussions of harnessing AI for a military advantage. However, assessment is a best practice we observed in historical examples of successful military innovation. Assessment was a critical component of rapid and effective learning. For example, in the adoption of helicopters in amphibious warfare, the first operational squadron worked closely with the doctrine command to perform experiments and

[39] Ilachinski, Andrew, AI, Robots, and Swarms: Issues, Questions, and Recommended Studies, CNA, January 2017.

assessments for concept development. Similarly, for carrier aviation, Admiral Joseph Mason Reeves, during his time at the Naval War College, conducted experiments to determine ways to improve carrier aviation, relying on detailed assessments to identify limiting factors and ways to address those limitations. Likewise, the Marine Corps development of maneuver warfare began with Major General Gray, the commander of 2nd Marine Division, directing a series of experiments that featured assessments to determine which tactics, technique, and procedures were effective. These assessments were not pass/fail determination but rather rigorous processes used to evaluate valuable components of tactical and operational approaches, which could then inform the refinement of operational concepts and, eventually, doctrine.

History shows that a rigorous assessment process can evaluate promising operational approaches, identify requirements, and reveal gaps and opportunities that can be further explored in future experiments, creating a learning loop. Thus assessments, if structured properly and resourced, can inform technical capabilities, current doctrine and training, and future experimentation events.

One area where assessments would be particularly valuable is exploring the nature of optimal human-machine teaming. With so many potential variables regarding the nature of automated functions, the roles of the human in a particular mission, and the ways that man and machine partner, rigorous assessments of that teaming and introducing an iterative process to accelerate learning would help the Navy more quickly leverage the potential benefits of human-machine teaming, while also rigorously assessing safety issues and the level of appropriate trust that is warranted in specific applications.

Such assessments do not have to be Navy-wide events. Individual commands can hold experimentation venues and perform assessments, or have external assessments performed in their support. The Navy could also seek joint opportunities for assessment resources, such as a Joint Test and Evaluation activity, which uses OSD funds to run evaluations and assessments.

Conclusions and Recommendations

Conclusions

AI is a powerful technology that has been likened to the third revolution in warfare, following the invention of gunpowder and nuclear weapons. Thus, AI carries many opportunities for the Navy to gain a military advantage. At the same time, the unique characteristics of AI also carry their own safety risks, and there are strategic reasons for the Navy to prioritize AI safety.

In this report, we have identified, analyzed, and suggested ways to address the primary safety risks associated with the Navy's use of AI. This work can serve as a starting point for the Navy to address more comprehensively the issues of AI safety that will arise as it adopts the technology. We also point out specific activities that the Navy can perform to manage AI safety risks, some in collaboration with academia and industry and others under the purview of the Navy.

We observe that seeking AI safety along with effectiveness is not a contradiction; the management of safety risks and achieving effectiveness can go hand in hand, as we saw in practice with SOF in recent operations. The primary challenge for the Navy in leveraging AI effectively and safely is not a technological one; rather, it begins with leadership. In historical examples of successful innovation, the key determiner for success began with leadership focus. We used historical examples of military innovation to point out some best practices the Navy can follow in its implementation of AI, including how safety considerations should be integrated into those practices.

Overall, it is true that the Navy is not in any danger of creating the Terminator or ending the world through its use of AI. However, there are a number of opportunities to better promote safety in its use of AI, by deliberately managing risks, such as those identified in this report, and taking actions proactively to address them. This report gives the Navy a foundation for acting to develop AI safety.

Recommendations

- Given that AI is a technology poised to revolutionize warfare, Navy leadership should convene a board that looks at the potential applications, benefits, and risks of AI, challenges common assumptions about the future operating environment and threats, and develops recommendations regarding a set of potential operating concepts to explore and pursue.

- The Navy should conduct regular experiments to test and explore new operating concepts designed to leverage AI. These experiments should include a focus on exploring concepts for human-machine teaming and autonomous systems.
- These experiments should be coupled with the refinement of operating concepts to help the Navy adapt to a new way of warfare, providing feedback to help the Navy leverage the strengths of AI. These operating concepts should promote effectiveness and address safety issues discussed in this report.
- The Navy should develop a rigorous assessment process to enable effective learning from experiments, provide feedback to concept development, and inform capability requirements. These assessments should pay close attention to identifying ways to optimize the human-machine team and cultivate appropriate trust, including ways to manage safety risks introduced by the use of AI.
- The Navy should work collaboratively with academia and with industry on the identified general safety risks of AI: fairness and bias, unpredictability and unexplainability, and cyber security and tampering. This collaboration can strengthen the safety enterprise of the Navy, including setting system requirements for safety, improving test and evaluation processes, refining the conduct of legal reviews, and, for autonomous systems, informing senior level reviews as required for DOD Directive 3000.09.
- The Navy should designate a point of responsibility and take a deliberate a risk management approach for AI applications, including efforts to establish context, identify risks associated with AI technology in general and specific planned applications of AI, analyze and evaluate those risks to inform their scope and severity, and then determine ways to address them.

Abbreviations

AI	Artificial Intelligence
C2	Command and Control
DIB	Defense Innovation Board
DOD	Department of Defense
DOJ	Department of Justice
IFF	Identification Friend or Foe
PPG	Presidential Policy Guidance
R&D	Research and Development
SOF	Special Operations Forces
T&E	Test and Evaluation

References

–, Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I), of 8 June 1977.

–, SKYNET, Terminator Wiki, https://terminator.fandom.com/wiki/Skynet

Allen, Gregory C., Understanding China's AI Strategy, CNAS, February 6 2019. https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research, 1-42, 2018.

Chartered Accountants, Establish the Context: Risk Management, https://survey.charteredaccountantsanz.com/risk_management/small-firms/context.aspx

Clark, Colin, Our Artificial Intelligence 'Sputnik Moment' is Now: Eric Schmidt & Bob Work, Breaking Defense, November 1, 2017, https://breakingdefense.com/2017/11/our-artificialintelligence-sputnik-moment-is-now-eric-schmidt-bob-work/.

Comey, James B., Hard Truths: Law Enforcement and Race, remarks delivered at Georgetown University, February 12 2015.

Defense Science Board, Report of the Defense Science Board Summer Study on Autonomy, Washington, DC: Office of the Secretary of Defense, June 2016, https://www.hsdl.org/?view&did=794641.

Defense Science Board, Summer Study on Autonomy, Department of Defense, June 2016.

Department of Defense, Summary of the 2018 Department of Defense Artificial Intelligence Strategy, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>.

Freedberg Jr., Sydney J., "Centaur Army: Bob Work, Robotics, and the Third Offset Strategy," Breaking Defense, November 9, 2015, https://breakingdefense.com/2015/11/centaur-army-bob-work-robotics-the-third-offset-strategy/

Griffith, Tom, Bias is Always Bad (answering What Scientific Idea is Ready for Retirement?), The Edge, 2014, https://www.edge.org/response-detail/25491.

Griggs, Mary Beth, People Trust Robots to Lead Them Out Of Danger, Even When They Shouldn't, Popular Science, March 1 2016, https://www.popsci.com/people-trust-robots-to-lead-them-out-danger-even-when-they-shouldnt/.

Hartnett, Kevin, Foundations Built for a General Theory of Neural Networks, Quanta Magazine, January 31 2019.

Hawking, Stephen, Stuart Russell, Max Tegmark, and Frank Wilczek, Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?', The Independent, May 1 2014, https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

Ilachinski, Andrew, AI, Robots, and Swarms: Issues, Questions, and Recommended Studies, CNA, January 2017.

International Standards Organization, Risk Management, ISO 51000, 2018, https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100426.pdf

Keller, John, Iran-U.S. RQ-170 incident has defense industry saying 'never again' to unmanned vehicle hacking, Military and Aerospace Electronics, May 2 2016.

Lewis, Larry, AI and Autonomy in War: Understanding and Mitigating Risks, CNA, August 2018.

Lewis, Larry, and Diane Vavrichek, An AI Framework for the Department of the Navy, CNA, August 2019.

Lewis, Larry, Insights for the Third Offset: Addressing Challenges of Autonomy and Artificial Intelligence in Military Operations, September 2017.

Lewis, Larry, Resolving the Battle over Artificial Intelligence in War, RUSI Journal, pre-publication draft, September 10 2019

Pellerin, Cheryl, Work: Human-Machine Teaming Represents Defense Technology Future, DEFENSE.GOV, November 8 2015.

Pierce, Terry C., Warfighting and Disruptive Technologies: Disguising Innovation, Frank Cass, 2004.

Powell, Alexander, Larry Lewis, Catherine Norman, and Jerry Meyerle, Summary Report: U.S.-UK Integration in Helmand, CNA, February 2016.

Ryan, Mick, Human-Machine Teaming for Future Ground Forces, Center for Strategic and Budgetary Assessments, April 2018.

Starry, Donn, To Change an Army, Military Review, 63 No. 3, March 1983

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, Intriguing properties of neural networks, ARXIV, 2013, https://www.arxiv-vanity.com/papers/1312.6199/.

The White House, 'Executive Order on Maintaining American Leadership in Artificial Intelligence', 11 February 2019, <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/;

The White House, Executive Order 13859, Executive Order on Maintaining American Leadership in Artificial Intelligence, February 11, 2019.

Thompson, Cadie, "Elon Musk Just Issued a Nightmarish Warning About What Will Really Happen if AI Takes Over," Science Alert, April 6, 2018, https://www.sciencealert.com/elon-musk-warns-that-creation-of-god-like-ai-could-doom-us-all-to-an-eternity-of-robot-dictatorship

Vincent, James, Putin says the nation that leads in AI 'will be the ruler of the world,' The Verge, September 4 2017, https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

Wafa, Abdul Waheed, and John F. Burns, U.S. Airstrike Reported to Hit Afghan Wedding, New York Times, November 5 2019, https://www.nytimes.com/2008/11/06/world/asia/06afghan.html.

Ziezulewicz, Geoff, The Ghost in the Fitz's Machine: why a doomed warship's crew never saw the vessel that hit it, Navy Times, January 14 2019.

This report was written by CNA's Operational Warfighting Division (OPS).

OPS focuses on ensuring that US military forces are able to compete and win against the nation's most capable adversaries. The major functional components of OPS work include activities associated with generating and then employing the force. *Force generation* addresses how forces and commands are organized, trained, scheduled, and deployed. *Force employment* encompasses concepts for how capabilities are arrayed, protected, and sustained at the operational level in peacetime and conflict, in all domains, against different types of adversaries, and under varied geographic and environmental conditions.

DOP-2019-U-021957-Final-2

3003 Washington Boulevard, Arlington, VA 22201

www.cna.org ● 703-824-2000