REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
05-06-2019	FINAL	July 2018 - June 2019

4. TITLE AND SUBTITLE

Trust in the Machine:
AI, Autonomy, and Military Decision Making with Lethal Consequences

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

CDR Christi S. Montgomery

Paper Advisor (if Any): CDR Michael O'Hara, PhD

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
N/A

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

Ethics and Emerging Military Technology Certificate Program
Naval War College
686 Cushing Road
Newport, RI 02841-1207

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSOR/MONITOR'S ACRONYM(S)

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

N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; Distribution is unlimited.

13. SUPPLEMENTARY NOTES Submitted to the Faculty of the U.S. Naval War College Newport, RI in partial satisfaction of the requirements of the Certificate Program in Ethics and Emerging Military Technology (EEMT).

14. ABSTRACT
The U.S. military has employed artificially intelligent and autonomous-capable weapons systems since the 1980s, but technology and capabilities have drastically changed in the past three decades. In order to remain competitive, the U.S. military leaders must reconsider AI and autonomous weapons employment doctrine across the spectrum of conflict, as well as work to improve trust in AI and autonomous technology. As codified in Department of Defense Policy, the development of artificial intelligence (AI) and autonomous weapons in mostly peacetime conditions has favored policy-maker insistence that military leaders who employ such technology exercise appropriate levels of human judgment over the use of force. In their insistence on human judgment, policy-makers made outdated assumptions about the availability of time, and implied a trust in the supremacy of human judgment over machine performance. Technological developments in recent years have compressed reaction times. The time available to bring lethal force to bear has decreased while the amount of contextual information that enables decisions on the use of force has increased. At the same time, gray zone conflict activity is increasingly blurring the line between peacetime operations and warfare. U.S. military forces exerting forward, deterrent presence in areas prone to activity that is not in accordance with international norms or law are increasingly exposed to complex risk that may be misunderstood and lethally miscalculated.

When operating in the gray zone, where the distinction between peace and war blurs and where technology has compressed reaction times, servicemembers face a moral gray zone. In the moral gray zone, operators encounter a potential dilemma between the duty to abide by the principle of distinction in the law of armed conflict (LOAC) and the inherent right to self-defense. The ambiguity inherent in the operational gray zone contributes to a higher-than-average likelihood of human judgment failures in the accompanying moral gray zone. AI and autonomous technology have the potential to improve both the success of self-defensive actions and the adherence to LOAC – particularly in compressed timescales – but only if humans and organizations are able to establish trust in the machine operating intelligently and autonomously. Establishing trust requires that humans perceive machine actions as predictable, transparent, and traceable. Trust also requires understanding how judgments of accountability, morality, and ethics differ between machine and human. Addressing these considerations in the development of new AI and autonomous systems will be necessary to ensure that servicemember and societal trust in the Department of Defense (DoD) is preserved.

15. SUBJECT TERMS
artificial intelligence, autonomous, autonomy, trust, decision-making, heuristics, gray zone, command, aegis, aegis weapons system, patriot missile system, ethics, morality, deontological, utilitarian, robot, machine, autonomous weapons systems, accountability, predictability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Director, EEMT Program
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED		59	401-841-7542

Standard Form 298 (Rev. 8-98)

Trust in the Machine:

AI, Autonomy, and Military Decision Making with Lethal Consequences

Christi S. Montgomery
CDR, United States Navy

Date Submitted: 05 JUN 2019

Submitted to the Faculty of the U.S. Naval War College Newport, RI in partial satisfaction of the requirements of the Certificate Program in Ethics and Emerging Military Technology (EEMT).

Thomas E. Creely, Ph.D. Date	Timothy Schultz, Ph.D. Date
EEMT Director	Associate Dean of Academics

CDR Michael P. O'Hara, Ph.D. Date
EEMT Mentor

Abstract

The U.S. military has employed artificially intelligent and autonomous-capable weapons systems since the 1980s, but technology and capabilities have drastically changed in the past three decades. In order to remain competitive, the U.S. military leaders must reconsider AI and autonomous weapons employment doctrine across the spectrum of conflict, as well as work to improve trust in AI and autonomous technology. As codified in Department of Defense Policy, the development of artificial intelligence (AI) and autonomous weapons in mostly peacetime conditions has favored policy-maker insistence that military leaders who employ such technology exercise appropriate levels of human judgment over the use of force. In their insistence on human judgment, policy-makers made outdated assumptions about the availability of time, and implied a trust in the supremacy of human judgment over machine performance. Technological developments in recent years have compressed reaction times. The time available to bring lethal force to bear has decreased while the amount of contextual information that enables decisions on the use of force has increased. At the same time, gray zone conflict activity is increasingly blurring the line between peacetime operations and warfare. U.S. military forces exerting forward, deterrent presence in areas prone to activity that is not in accordance with international norms or law are increasingly exposed to complex risk that may be misunderstood and lethally miscalculated.

When operating in the gray zone, where the distinction between peace and war blurs and where technology has compressed reaction times, servicemembers face a moral gray zone. In the moral gray zone, operators encounter a potential dilemma between the duty to abide by the principle of distinction in the law of armed conflict (LOAC) and the inherent right to self-defense. The ambiguity inherent in the operational gray zone contributes to a higher-than-average likelihood of human judgment failures in the accompanying moral gray zone. AI and autonomous technology have the potential to improve both the success of self-defensive actions and the adherence to LOAC – particularly in compressed timescales – but only if humans and organizations are able to establish trust in the machine operating intelligently and autonomously. Establishing trust requires that humans perceive machine actions as predictable, transparent, and traceable. Trust also requires understanding how judgments of accountability, morality, and ethics differ between machine and human. Addressing these considerations in the development of new AI and autonomous systems will be necessary to ensure that servicemember and societal trust in the Department of Defense (DoD) is preserved.

Table of Contents

Introduction

The Palawan coastline was fading quickly from navigational radar as USS NITZE made a course within 12 nautical miles of Fiery Cross Reef in the Spratly Islands to conduct a Freedom of Navigation Operation, or "FONOP" for short. The sun had slipped below the horizon, and there were only a few more minutes of nautical twilight. In the suspense between day and night, the bridge team's visual acuity ever so slightly decreased as their eyes adjusted,[1] and conventional infrared sensors on the ship lost thermal detection contrast from thermal crossover.[2] The Officer of the Deck stood on the bridge, binoculars glued to her eyes, visually and mentally charting the safest course through the congested and terrestrially punctuated waters. Piercing the silence, the Tactical Action Officer's voice, "*Bridge, Combat, we have MISSILES INBOUND bearing 310°!*" This message was followed seconds later with "*General Quarters, General Quarters. All hands, man your battle stations*," over the ship's main circuit (1MC). Immediately, the aft cells of the vertical launch system (VLS) violently expelled their loads of SM-2 and Evolved Sea Sparrow Missiles (ESSMs) to shoot down the incoming missiles, but just as the payloads began to climb, a salvo of YJ-12 anti-ship cruise missiles (ASCMs) made their supersonic descent into the hull of NITZE at Mach 3. By the time General Quarters was called, the ship had less than 5 seconds before the inbound missiles hit their target. Although designed for autonomous operation, the powerful Aegis Weapons System on NITZE had been configured for this peacetime mission to require operator permission before engagement. The key for the Fire Inhibit Switch (FIS) was turned to the off position, disabling the VLS,[3] and the Captain

---

[1] Marc Green, "Night Vision," *Marc Green PhD Human Factors*, accessed May 25, 2019, https://www.visualexpert.com/Resources/nightvision.html.

[2] Huijie Zhao et al., "Target Detection over the Diurnal Cycle Using a Multispectral Infrared Sensor," *Sensors* 17, no. 56 (Jan 2017): 3, accessed 25 May 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5298629/pdf/sensors-17-00056.pdf.

[3] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, 2018), 165.

employed the autonomous-capable Aegis system with an operator in the loop. This practice is standard operating procedure – a product of a peacetime Navy long accustomed to "disciplined restraint."[4] The Aegis Weapons System was designed to save the ship in moments like this. However, as a salvo of supersonic missiles bee-lined for them, the slow reaction time of the humans-in-the-loop doomed their ship to destruction.

This fictional account may seem far-fetched and remind you of the scenarios described with prescient clarity in the popular novel, *Ghost Fleet*.[5] Unfortunately, the likelihood of the above vignette becoming reality is increasing by the day. China has installed supersonic YJ-12B ASCMs on Fiery Cross Reef,[6] and the U.S. Navy FONOP program routinely sends Aegis destroyers in to the contested South China Sea.[7] The FONOP program referenced in the vignette occurs in an area of the South China Sea that has seen a rise in military and state activity that could be classified as 'gray zone' activity[8] – not quite peaceful, but also not quite warfare.[9] The political sensitivities of operating in the gray zone mean that tolerance for error is extremely low. An accidental misfire of any weapon from a U.S. Navy ship, could signal to China that the U.S.

---

[4] David Crist, *The Twilight War: The Secret History of America's 30-Year Conflict with Iran* (New York: Penguin Books, 2012), 560. Though this fictional vignette occurs in the South China Sea, the moniker for the naval rules of engagement in the 5th Fleet AOR from 2008-2011 was "disciplined restraint." The author experienced the tension contained in this ROE as a member of the bridge watch team aboard USS THEODORE ROOSEVELT during multiple transits of the Straits of Hormuz. Iranian Revolutionary Guard speedboats commonly harassed and drove threatening lines of approach to the U.S. aircraft carrier and escort ships. Though called something different in 7th Fleet, the 'hold-your-fire' approach to operating in contested waterways which characterized disciplined restraint is still present.

[5] Peter Singer and August Cole, *Ghost Fleet* (New York: Houghton Mifflin Harcourt Publishing, 2015).

[6] David Brunnstrom, "China Installs Cruise Missiles on South China Sea Outpost: CNBC," *Reuters*, May 2, 2018, https://www.reuters.com/article/us-southchinasea-china-missiles/china-installs-cruise-missiles-on-south-china-sea-outposts-cnbc-idUSKBN1I336G.

[7] Eleanor Freund, "Freedom of Navigation in the South China Sea: A Practical Guide," Belfer Center for Science and International Affairs, Harvard Kennedy School, last modified June 2017, https://www.belfercenter.org/publication/freedom-navigation-south-china-sea-practical-guide

[8] Michael Green et al., "Coercion in Maritime Asia: The Theory and Practice of Gray Zone Deterrence," (Washington, DC: Center for Strategic and International Studies (CSIS), May 2017): 3, https://csis-prod.s3.amazonaws.com/s3fs-public/publication/170505_GreenM_CounteringCoercionAsia_Web.pdf?OnoJXfWb4A5gw_n6G.8azgEd8zRIM4wq.

[9] Green et al., "Coercion," 21-22.

is attacking its sovereignty, which has the potential to escalate tensions between the U.S. and China and increase the risk of conflict.[10] U.S. FONOPs are intended to "send a general message of U.S. resolve, as well as demonstrate that Washington [will] not recognize any illegal Chinese claims to additional maritime rights based on the artificial expansion of its occupied features."[11] Though the U.S. attempts to exert extraordinary human discipline in the execution of FONOPs, these activities repeatedly incite Chinese condemnation and public outrage.[12] A FONOP within 12 nautical miles of Fiery Cross Reef by USS WILLIAM P. LAWRENCE in 2016 elicited an immediate response by three Chinese warships and two fighter jets.[13] Though China has not given indications or warnings that it will resort to the use of lethal force to counter U.S. deterrence, the demonstrated aggression of their military forces yields a greater-than-zero chance that the Chinese might shoot first.[14]

Into this milieu, the U.S. military employs artificially intelligent, autonomous-capable weapons, and has since the 1980s. But technology and capabilities have drastically changed in the past three decades. In order to remain competitive, the U.S. military leaders must reconsider AI and autonomous weapons employment doctrine across the spectrum of conflict, as well as work to improve trust in AI and autonomous technology. The development of artificial intelligence (AI) and autonomous weapons in mostly peacetime conditions has favored policy-maker insistence that military leaders who employ such technology exercise "appropriate levels of human judgment over the use of force." In their insistence on human judgment, policy-

[10] Green et al., "Coercion," 4.
[11] Green et al., 248.
[12] Green et al., 253.
[13] Green et al., 254.
[14] Graham Allison, The Thucydides Trap: Are the U.S. and China Headed for War?" The Atlantic, September 24, 2015, https://www.theatlantic.com/international/archive/2015/09/united-states-china-war-thucydides-trap/406756/.

makers have made erroneous assumptions about the availability of time, and implied a trust in the supremacy of human judgment over machine performance.

Technological developments in recent years have compressed reaction times. The time available to bring lethal force to bear has decreased while the amount of contextual information that enables decisions on the use of force has increased. At the same time, gray zone conflict activity is increasingly blurring the line between peacetime operations and warfare. U.S. military forces exerting forward, deterrent presence in areas prone to activity that is not in accordance with international norms or law are increasingly exposed to risk that may be misunderstood and lethally miscalculated. Critically, training and doctrine for peacetime use of AI and autonomy, where servicemembers are expected to trust human judgment over machine, is different than training and doctrine for wartime use of AI and autonomy, when the expectation of violence requires trust in the machine for survival.

When operating in the gray zone, where the distinction between peace and war blurs and where technology has compressed reaction times, servicemembers face a *moral* gray zone. In the moral gray zone, operators encounter a potential dilemma between the duty to abide by the principle of distinction in the law of armed conflict (LOAC) and the inherent right to self-defense. The ambiguity inherent in the operational gray zone contributes to a higher-than-average likelihood of human judgment failures in the accompanying moral gray zone. AI and autonomous technology have the potential to improve both the success of self-defensive actions and the adherence to LOAC – particularly in compressed timescales – but only if humans and organizations are able to establish trust in the machine operating intelligently and autonomously. Establishing trust requires that humans perceive machine actions as predictable; that humans have knowledge of how the machine makes decisions through transparency and traceability; and

that humans understand how judgments of accountability and morality may be different for a machine than for a human. Addressing these considerations in the development of new AI and autonomous systems for military use will be necessary to ensure that servicemember and societal trust in the Department of Defense (DoD) is preserved.

AI and Autonomy in the Department of Defense – Now and in the Future

A framework for an expanded exploration of AI and autonomous technology applications in the military starts with a clear understanding of both Artificial Intelligence and Autonomy. The term 'Artificial Intelligence' was first coined in 1956 by a group of researchers at a summer workshop on the topic at Dartmouth College.[15] Since then, the term has expanded in conception and utility, to encompass a broad swath of technological and innovative development, especially in the fields of computer science and robotics. However, as the decades have passed, agreement on a definition of the term has become increasingly difficult. Though 'artificial' seems easy enough to conceptualize, defining 'intelligence' has been fodder for millennia of philosophers and scientists alike. In Plato's *Theaetetus,* Socrates circumscribes intelligence in terms of 'knowledge,' and knowledge in terms of 'perception,' 'arts and sciences,' 'true judgment,' and 'true judgment with logos [logic].' Socrates highlights Protagoras' description of knowledge as belonging to man, "man is the measure of all things, of the existence of the things that are and the non-existence of the things that are not."[16] If intelligence is a derivation of human perception and judgment, a derivative definition of artificial intelligence might be something non-human (artificial) that can possess human perception and judgment. Socrates also emphasizes the aspect

[15] McCarthy et al, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," (unpublished research proposal, Dartmouth College, August 31, 1955), 1, accessed 11 Apr 2019, http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

[16] Plato, "Theaetetus," *Plato in Twelve Volumes*, trans. Harold N. Fowler (Cambridge, MA: Harvard University Press, 1921), accessed May 25, 2019, http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0172%3Atext%3DTheaet.%3Apage%3D152.

of motion, or action, as an aspect of intelligence.[17] This introduces an element of action to the definition.

Accordingly, modern understanding of artificial intelligence is often classified along two dimensions, thinking and acting,[18] and is further divided between thinking and acting humanly, or thinking and acting rationally.[19] Peter Norvig and Stuart J. Russell utilize these categories to bin the various types of AI in *Artificial Intelligence: A Modern Approach*,[20] which is "used to teach AI researchers around the world."[21] Max Tegmark's simpler definition of artificial intelligence, which is the "non-biological ability to accomplish complex goals,"[22] is most useful because it opens conceptual thinking about what may be deemed AI. A widened aperture on AI, particularly within the DoD, supports the assertion that the U.S. military has been employing AI for decades.

Within most definitions of AI one can find a subset of classifications - narrow AI or general AI. Narrow AI has the "ability to accomplish a narrow set of goals."[23] It is generally applied to accomplishing goals with bounded, or finite, solutions or strategies. An example is the game of chess. Successfully defeating the human grand champion chess master may seem complex, but there is a finite set of moves in chess, and it is a game well suited for narrow AI. General AI theoretically has the "ability to accomplish virtually any goal, including learning,"[24]

---

[17] Plato, "Theaetetus," Plato in Twelve Volumes, trans. Harold N. Fowler (Cambridge, MA: Harvard University Press, 1921), accessed May 31, 2019, http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0172%3Atext%3DTheaet.%3Apage%3D153.
[18] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, (London: Pearson Education 2010), 1.
[19] Russell and Norvig, *Artificial*, 2.
[20] Russell and Norvig, 2.
[21] Scharre, *Army of None*, 68.
[22] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Vintage Books, 2017), 39.
[23] Tegmark, *Life 3.0*, 39.
[24] Tegmark, 39.

and is usually described as human-level intelligence. The term 'theoretical' is used because general AI has never been developed. Not only have scientists not come close to figuring out the best way to build general AI, the limitations of today's computer hardware preclude it. For the purposes of this paper, all use of the term artificial intelligence refers to narrow AI. This does not exclude the AI with the ability to learn, but it does limit consideration to AI with specific, achievable goals.

A goal-oriented perspective of AI is useful in understanding how AI and autonomy may complement one another. Autonomy is the term used to describe both the level of human involvement with a machine's ability to accomplish a goal, and the complexity of the machine/system's decision making.[25] The umbrella of autonomy extends over a range of machine intelligence. Machines that "sense the environment and act" (e.g. your thermostat sensing the temp and turning on or off the furnace)[26] are categorized as *automatic*. These systems are simple to understand and ubiquitous. Machines that are complex, but still rule-based,[27] are categorized as *automated* (e.g. an Automated Teller Machine). Though complex, the actions of automated systems are more or less traceable. Machines that are highly complex, "goal-oriented and self-directed" [28] are categorized as *autonomous*. For cases of highly complex autonomous systems, there are different levels of human intervention or interaction.

The level of human intervention is often described by the terms: human-in-the-loop or *semi-autonomous*, human-on-the-loop or *supervised-autonomous*, and human-out-of-the-loop or *fully-autonomous*.[29] In semi-autonomous/human-in-the-loop systems, the human "must remain

---

[25] Scharre, *Army of None*, 27.
[26] Scharre,, 30.
[27] Scharre, 31.
[28] Scharre, 31.
[29] Scharre, 29-30. Though Scharre's definitions of autonomy are used here, they mirror the definitions included in DoD Directive (DODD) 3000.09, *Autonomy in Weapon Systems*, (Washington, DC: Department of Defense, 2012), 13-14, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. The definition of

[…] an active participant,"[30] in the accomplishment of the goal. Often, that involves the setting of the goal, and the determination of actions. The Mars Curiosity Rover is an example of a semi-autonomous machine. In human-on-the-loop/supervised-autonomous systems, the system retains the ability to sense, plan, and act independently, but the human monitors the system's achievement of goals and can interrupt the machine if there are any issues.[31] Automobile assembly-line robots are examples of supervised autonomous machines. In fully-autonomous systems, the machine "performs all aspects of a task autonomously without human intervention with sensing, planning, or implementing action."[32] These terms and categories, while simple, reflect a common understanding of machine autonomy in human-machine (or human-robot) interaction.[33]

Autonomous/semi-autonomous weapons are sub-categories of AI and are instantiations of intelligent agents. The definitions of AI and autonomy are sometimes conflated and misunderstood, including within the DoD, where resourcing for AI and autonomous systems are often programmatically separate and governed by separate policy. In 2012, DoD published Department of Defense Directive (DoDD) 3000.09, *Autonomy in Weapon Systems*, which "establishes DoD policy and assigns responsibilities for the development and use of autonomous and semi-autonomous functions in weapon systems, including manned and unmanned

---

'Autonomous Weapon System' in DODD 3000.09 is a "weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation." A 'Semi-Autonomous Weapon System' is a "weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator."

[30] William D. Nothwang et al., "The Human Should be Part of the Control Loop?," (unpublished research, Office of the Secretary of Defense Autonomy Research Pilot Initiative, 2016), 1, accessed 29 May 2019, http://faculty.washington.edu/sburden/_papers/NothwangRobinson2016resil.pdf.

[31] Jenay Beer, A.D. Fisk, and W.A. Rogers, "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction* 3, no. 2 (2014), 85-87, https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1127&context=csce_facpub.

[32] Beer, Fisk, and Rogers, "Toward a Framework," 87.

[33] Beer, Fisk, and Rogers, 85-92,

platforms."[34] While the DoD had several decades of experience with the use of autonomous or semi-autonomous weapons, a complex understanding of AI – including past and present use and opportunities for the future – was still relatively new in military and policy circles in 2012. Published in February 2019, the DoD AI strategy provides limited reference to a definition of AI in the following, "AI refers to the ability of machines to perform tasks that normally require human intelligence – for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action – whether digitally or as the smart software behind autonomous physical systems."[35] The continued perpetuation of 'autonomous weapons' as a separate concept may be an artifact from a limited conceptual understanding of the breadth and scope of the artificial intelligence field. The DoD AI Strategy definition of AI may skew the focus of the strategy to human-level intelligence, which is highly aspirational, and may cause researchers and policy makers to ignore applicable lessons and take-aways from decades of AI application.

Consistent throughout strategy and policy on both AI and autonomy is the notion of leveraging appropriate human judgment as a safeguard for lethality and accountability. Autonomous weapons policy very clearly spells out a requirement for human judgment – or human-in-the-loop/on-the-loop – design. DoDD 3000.09 specifies that, "autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise *appropriate* [emphasis added] levels of human judgment over the use of force."[36] The directive also requires all organizations in the DoD to, "design autonomous and semi-autonomous weapon

---

[34] Department of Defense, *Autonomy in Weapon Systems*, DoD Directive (DODD) 3000.09 (Washington, DC: Department of Defense, 2012), 1, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.
[35] Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance our Security and Prosperity* (Washington, DC: Department of Defense, 2018), 5, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
[36] Department of Defense, *Autonomy in Weapon Systems*, 2.

systems in such a manner as to minimize the probability and consequences of failures that could lead to unintended engagements or to loss of control of the system."[37] Though it references the above guidance on autonomous development as a principle of ethics and safety, the DoD AI Strategy contains no language regarding the requirement for relative engagement of a human with a military-purposed artificially intelligent agent. There is no insistence on a human-in or on-the-loop in the DoD AI Strategy. This either conforms to the continued separation of AI and autonomy in conceptual application, or leaves an opening for AI applications with no direct human oversight. In both documents, what is clear is an insistence that autonomous weapons and artificial intelligence are used consistent with "the law of war and our nation's values."[38] What is still unclear is a definition of *appropriate* levels of human judgment.

The rapid pace of technological developments during this decade suggest the successful military integration of artificial intelligence and autonomous weapons is a new field of exploration, but military development of AI-enabled, autonomous-capable weapons systems happened as early as the late 1960s. The Aegis Weapons System and the Patriot Missile System are both examples of narrow-AI enabled, autonomous-capable weapons systems, designed in the late 60s and employed by the U.S. military today. A review of their use provides extraordinary insight into the operational employment of such systems, across multiple decades and threat scenarios, with important human-machine interface lessons that can be applied to the current and future development of military AI and autonomous weapons technology. The review of human perception after years of interfacing with these systems is critical to understanding how human trust in AI and autonomy may, or may not, evolve to meet the complex combat challenges of the future, and may help DoD to better define appropriate levels of human judgment. The following

[37] Department of Defense, *Autonomy in Weapon Systems*, 11.
[38] Department of Defense, *Summary of the 2018*, 15.

cases will demonstrate that the human-in-the-loop should not, and cannot, be the backstop to mitigate uncertainty in AI and autonomous technology performance, and that policy and strategy which place this burden on commanders and command decision-making may be lethally overestimating the reliability of human judgment. They will also show that the level of trust in AI and autonomy is a product of predictability, knowability, morality, and accountability – all factors that vary dynamically across the spectrum of conflict from peace to war.

The Aegis Weapons System – Exploring Existing Narrow AI and Autonomy

The Aegis Weapons System (AWS), also known as the Aegis Combat System, was designed in the late 1960s and developed in "the 1970s for defending ships against aircraft, anti-ship cruise missiles (ASCMs), surface threats, and subsurface threats."[39] It is a complex system built of nine interfacing components. The centerpiece is the AN/SPY fixed and phased-array radar. This is the primary sensor, which has the ability to "perform search, track and missile guidance functions simultaneously, with a track capacity of more than 100 targets."[40] AWS was "designed as a total weapon system, from detection to kill."[41] The brains of the system reside in the Command and Decision element (C&D) and Weapons Control System (WCS).[42] The C&D element allows the human operator to program automatic functionality into the AWS through "control by doctrine,"[43]which consists of a number of conditional logic, or 'if-then' statements regarding the assessment of a specific radar-selected track, that a human operator uploads, at a

---

[39] Ronald O'Rourke, *Navy Aegis Ballistic Missile Defense (BMD) Program: Background and Issues for Congress* (Washington, DC: Congressional Research Service, 2019),1, https://crsreports.congress.gov/product/pdf/RL/RL33745.
[40] U.S. Navy, "U.S. Navy Fact File: AEGIS Weapons System," accessed April 9, 2019, https://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2
[41] U.S. Navy, "Navy Fact File."
[42] "AEGIS Weapons System Mk 7," Global Security, accessed April 22, 2019, https://www.globalsecurity.org/military/systems/ship/systems/aegis-core.htm.
[43] John R. Gersh, "Doctrinal Automation in Naval Combat Systems: The Experience and the Future," *Naval Engineers Journal* 99, no. 3 (May 1987): 74.

computer console into the C&D element. These conditional logic statements are called "doctrine statements,"[44] and they classify radar-selected tracks by the characteristics of "geometry (range, bearing, altitude, x-y coordinates), kinematics (course, speed, inbound/outbound, CPA), identity, Identification-Friendly-or-Foe (IFF) response, and category (air, surface, subsurface)."[45]

Prior to a deployment, an Aegis-capable ship's crew obtains the parameters and characteristics of likely threats within an area of responsibility (AOR). The Aegis operators are responsible for writing general, and threat-specific doctrine statements which provide the AN/SPY radar with sectors for concentrated focus. In this manner, the operators prime the system by providing it with non-algebraic, 360°, 0-90° azimuth, probability-of-detection estimates. Doctrine statements should be updated when the ship transits to another AOR, in recognition of geographically specific threats. When the radar receives a return from a contact, the system immediately begins the process of classifying the radar return as a track using the above-mentioned characteristics of category, geometry, kinematics, identity, and IFF response. It might determine that the speed, altitude, bearing and range of a track, coupled with military Mode II IFF signal, indicate the track is a U.S. military aircraft. It should be able to distinguish between the U.S. military aircraft and an incoming missile based on the above listed parameters. If the track characteristics indicate a possible hostile threat – like an incoming missile - the system will derive a fire control solution, and can – without human intervention – proceed to selecting an effective weapon response. The set of possible responses includes launching a missile to intercept, or engage, the incoming target.

But the U.S. Navy does not operate AWS without a human in the process. Throughout the entire process of detection, classification, identification, weapon selection, and engagement a

[44] Gersh, "Doctrinal Automation," 74.
[45] Gersh, "Doctrinal Automation," 75.

human is monitoring the system using various interactive consoles and displays on the ship. The AWS has the capability to operate in this supervised-autonomous manner with a human-on-the-loop, [46] or monitoring without direct input in to the process. However, U.S. Navy surface doctrine and operational use has put a deliberate human break before engagement in the detect-to-engage sequence of the AWS. This break is so deliberate that it is not only built into the doctrine statements which guide the performance of Aegis, but it also involves a keyed, analogue switch, called a Fire Inhibit Switch (FIS), which enables/disables the missile vertical launching mechanisms. This human break effectively reduces the system's autonomy from supervised autonomous (human-*on*-the-loop) to semi-autonomous (human-*in*-the-loop).

Interviews with combat systems operators from Aegis-capable ships presented two interesting perspectives regarding Aegis operation in a semi-autonomous versus supervised-autonomous manner. First, the decision to use lethal force was inherent in the responsibility of Command. Although the Commanding Officer may delegate some responsibilities to tactical action officers on U.S. Navy ships, the normative understanding is that the decision for lethal engagement should always be made by a human – and by the Commanding Officer. Second, there is a lack of trust in the AWS's capability to perform to the standard required or desired (yet undefined) for trusting it in a supervised-autonomous mode.[47] Much of the deficit in confidence stems from both a lack of understanding of how the doctrine statements interact in the machine logic of Aegis, and the basic mechanics and fallibility of a radar system operating in dynamic atmospheric conditions. A dynamic, turbulent atmosphere and clutter-inducing coastal

[46] Scharre, *Army of None*, 45.
[47] CDR Joe McGettigan, interview by author, 24 April 2019. CDR McGettigan served as the Air Defense Instructor at the U.S. Navy Surface Warfare Officer School (SWOS) from 2016 to 2018, and has served as the Combat Systems Officer on an Aegis-capable, U.S. Guided Missile Destroyer.

geography[48] can cause any surface-search radar system to either 'sense' contacts that do not exist, or fail to 'sense' contacts that do exist. The contact detection error rate of the AN/SPY radar is classified. However, it is not fallacious to suspect that the error rate is significant enough to induce a lack of confidence in operators, who ostensibly hold a standard that requires AWS to accurately detect all real contacts, and disregard all false ones. Keeping a human in the loop and depending on human judgment to compensate for machine error are safeguards for low trust and confidence in the performance of the system. Continuing to employ human back-stops for low-trust systems becomes increasingly dangerous as great-power competitors develop, deploy, and proliferate supersonic anti-ship missiles, and the time for action between detection and engagement is a miniscule fraction of what it has been in the past, and certainly not on the "organic timetable"[49] of human-fought warfare.

Patriot Missile System – Narrow AI and Human-on-the-Loop

A similar supervised autonomous weapons system in use for ground-based, air-defense by the U.S. military today got its start at the same time as the Aegis Weapons System. The DoD initiated a joint investigation in the late 1960s to determine if the Army's development of the Mobile Field Army Air Defense System could be combined with the Navy's development of the Advanced Surface Missile System (ASMS) – a precursor to AWS. It was determined that "complete commonality was not practical,"[50] and the two systems continued on divergent development paths. These divergent paths led to very different training, doctrine, and human-machine interface behavior for systems that are extraordinarily similar. The Army's system later

---

[48] "AN/SPY-1 Radar," Missile Defense Advocacy Alliance, December 2018, http://missiledefenseadvocacy.org/missile-defense-systems-2/missile-defense-systems/u-s-deployed-sensor-systems/anspy-1-radar/.

[49] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (New York: HarperCollins Publishers, 2017), 312.

[50] James D. Flanagan and William N. Sweet, "AEGIS: Advanced Surface Missile System," *Johns Hopkins APL Technical Digest* 2, no. 4 (1981), accessed May 18, 2019, https://www.jhuapl.edu/techdigest/views/pdfs/V02_N4_1981/V2_N4_1981_Flanagan_Advanced.pdf, 244.

became the Phased Array Tracking Radar to Intercept on Target (PATRIOT) Missile System. During the 1991 and 2003 wars in Iraq, the U.S. Army utilized Patriot Missile batteries for defense against the enemy use of ballistic missiles. The operational concept incorporates the detection of an incoming ballistic missile threat, the classification of the threat, and finally – the launch of the MIM-104 Patriot surface-to-air missile to intercept and eliminate the incoming ballistic missile. To get to a targeting solution, the Patriot's radar targeting "system applies complex computer algorithms to judge a target's speed and altitude and, in the case of an airplane, its radio transponder signal. If the computer decides a bogey matches the profile of an enemy aircraft or missile, it displays the target as hostile on operators' screens."[51] This process is nearly identical to the AWS process, though executed with different sensor and targeting components. Originally built in the 1970s as an anti-aircraft weapon, Patriot Missile batteries were in service operation in the mid-1980s as air-defense weapons.[52] They were touted for their use in defense against Iraqi Scud missiles in the 1991 Gulf War.

The Patriot's use in 2003 garnered significant attention over three incidents of friendly fire, two of which resulted in fratricide and coalition casualties. The Defense Science Board (DSB) reviewed the performance of the Patriot in the 2003 Operation Iraqi Freedom (OIF) in a report published in January 2005. The board identified three contributing causes to the friendly-fire incidents: 1) poor performance of the Mode IV Identification-Friendly-or-Foe (IFF) system, 2) lack of human situational awareness from a failure to integrate information into a combined air defense common operating picture, and 3) the adaptation of Patriot system tactics, operating

---

[51] David Axe, "That Time an Air Force F-16 and an Army Missile Battery Fought Each Other," *Medium*, July 5, 2014, https://medium.com/war-is-boring/that-time-an-air-force-f-16-and-an-army-missile-battery-fought-each-other-bb89d7d03b7d

[52] Charles Pillar, "Vaunted Patriot Missile has a "Friendly Fire" Failing," *L.A. Times*, April 21, 2003, https://www.latimes.com/archives/la-xpm-2003-apr-21-war-patriot21-story.html

procedures, and software algorithms for operation in Iraq.[53] Of particular concern was the "automatic" operating protocol, where "operators were trained to trust the system's software; a design that would be needed for heavy missile attacks."[54] In this wartime employment case, Patriot was operating in supervised autonomous mode, with humans-*on*-the-loop. The Patriot battery commander and operators were able to monitor the function of the Patriot system, and intercede if there was a problem, but generally allowed the system to sense, plan, and act without human input. In contrast to the U.S. Army's *expected* operating environment for employment of Patriot, the first "30 days of OIF involved nine engagements of tactical ballistic missiles which were immersed in an environment of some 41,000 coalition aircraft sorties; a 4,000-to-1 friendly-to-enemy ratio."[55] This amount of airspace congestion was overwhelming for humans and machines alike. In this wartime scenario, the human operators should trust the machine to outperform them in detecting, classifying, and engaging a potential threat. However, unknown to the human operators, their expectations of the machine in this environment exceeded the limits of its ability.

Use of the Patriot system in OIF in 2003 is an early example of the use of narrow artificial intelligence to make detection and engagement decisions, and the results have important implications for future warfighting. In one of the incidents sited in the DSB report, a U.S. Air Force F-16 actually fired upon and destroyed a Patriot Battery's radar system after being 'locked-on' by the Patriot's fire control system.[56] The engagement was classified as an accident, but a pilot interviewed afterwards anonymously shared sentiments of relief upon learning of the

---

[53] Defense Science Board, *Report of the Defense Science Board Task Force on Patriot System Performance – Report Summary* (Washington, DC: Office of the Under Secretary of Defense For Acquisition, Technology, and Logistics, January 2005), 2, accessed April 10, 2019, https://www.hsdl.org/?view&did=454598, 2.
[54] Defense Science Board, *Report of the Defense Science Board*, 2.
[55] Defense Science Board, 2.
[56] Axe, "That Time an Air Force F-16."

destruction stating, “no one was hurt when the Patriot was hit, thank God, but from our perspective they’re now down one radar. That’s one radar they can’t target us with any more.”[57] When the third friendly-fire incident resulted in the death of a Navy F/A-18 pilot, Patriot operators were instructed not to put the system on “fully-automatic modes.”[58] An embedded reporter, Robert Riggs, described his experience with a Patriot battery team, tracking air targets in Iraq:

> *This was like a bad science fiction movie in which the computer starts creating false targets. And you have the operators of the system wondering is this a figment of a computer's imagination or is this real. They were seeing what were called spurious targets that were identified as incoming tactical ballistic missiles. Sometimes, they didn't exist at all in time and space. Other times, they were identifying friendly U.S. aircraft as incoming TBMs.*[59]

All three incidents involve the Patriot’s radar tracking system – the AN/MPQ-53 "Phased Array Tracking Radar to Intercept on Target” – as reporting false or spurious targets, in addition to other system deficiencies. The radar misidentified coalition aircraft as incoming enemy missiles.[60] The recommendations of the DSB included shifting “its operation and control philosophy to deal with the complex environments of today’s and future conflicts. These future conflicts will likely be more stressing than OIF and involve Patriot in simultaneous missile and air defense engagements. A protocol that allows more operator oversight and control of major system actions will be needed.”[61] OIF Commanders, and subsequently the DSB, implemented a requirement for the human-*in*-the-loop for decisions regarding lethal engagement with the Patriot Missile System, even though it was being used in wartime, with very real, physical threat.

---

[57] Axe, “That Time an Air Force F-16.”
[58] Axe.
[59] Robert Riggs, “Embedded in Iraq with 5/52 ADA Patriot Missile Battalion,” quoted in Rebecca Leung, “The Patriot Flawed? Failure to Correct Problems Led to Friendly Fire Deaths,” *CBS News*, February 19, 2004, https://www.cbsnews.com/news/the-patriot-flawed-19-02-2004/.
[60] Robert Riggs, “Patriot Missile Friendly Fire Investigation,” *CBS 11 News*, Video, 15:35, June 2004, https://youtu.be/MugiYvjiOzA.
[61] Defense Science Board, *Report of the Defense Science Board*, 3.

Though employed by the U.S. military for the same purpose – air defense – comparing the doctrinal use of Aegis Weapons System and Patriot Missile System provides an interesting contrast in perception of threat and trust in the machine. In the 2003 use of the Patriot, the system was being employed in war, with a known and limited threat type. After the first incident of fratricide involving the use of the Patriot system, the Army opened an investigation, but continued to operate the system in a supervised autonomous mode because the threat environment had not changed, and the perceived risk of additional incidents was low.[62] The specificity of the Patriot mission, and the narrow geographic scope of employment, meant that Army doctrine called for supervised autonomous employment of the system, by personnel with limited experience and junior rank. The Aegis Weapons System has also been employed in war, but the system is customizable to a wider range of threat types, with a 360° engagement envelope. Because of the capability and lethality of the system, and the potential for unintended geostrategic implications of misfire, Navy doctrine requires semi-autonomous employment of the AWS. This employment paradigm, when compared to that of the Patriot Missile System, provides some insight into the differing doctrinal norms that have developed for the two systems. Though AWS has been employed in wartime, the Navy's legal and cultural practice of according ultimate trust, accountability, authority, and responsibility in the Commanding Officer has inculcated the use of a human-backstop for Aegis. This creates an unreasonable expectation of near-omniscience and non-human cognitive response by the CO in the maritime battlespace, where the threshold for what is *expected*, and what is *overwhelming*, is leveled much higher than that of a Patriot Missile System commander defending a ground-based target. In Figure 1 below, the answer to the question in block (A) for a Patriot Missile System commander is invariably

[62] Scharre, *Army of None*, 141.

"*Yes*." For a Navy ship commander with a much more capable Aegis Weapons System, the answer is almost always "*No*."

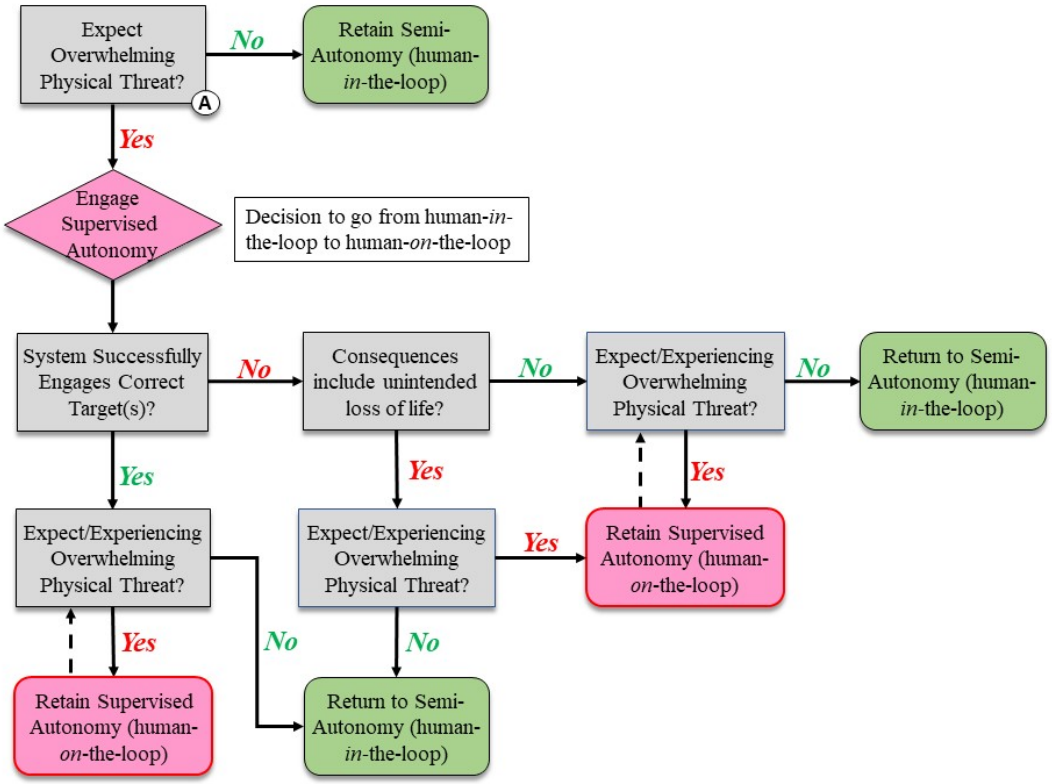

Figure 1. Flow chart for decisions regarding autonomous employment of weapons systems.

The Third Offset Strategy, AI, and Autonomy

The Aegis Weapons System and the Patriot Missile System were conceived, designed, and deployed during the Cold War. The DoD, haunted by the specter of the Soviet Union achieving nuclear and conventional parity, sought "technology investments in conventional forces…"[63] that "…could restore America's deterrence umbrella in Europe and offset the Soviet threat."[64] Past offset strategies have been frameworks for utilizing an asymmetric advantage to

[63] Robert Tomes, "The Cold War Offset Strategy: Origins and Relevance," War on the Rocks, November 6, 2014, https://warontherocks.com/2014/11/the-cold-war-offset-strategy-origins-and-relevance/.
[64] Tomes, "The Cold War."

dominate a great-power competitor. In 2014, the DoD published the "Third Offset Strategy." The Third Offset, similar to previous strategies, has "technological and operational innovation" as its central tenet, with five components: 1) Deep-Learning Systems, 2) Human-Machine Collaboration, 3) Human-Machine Combat Teaming, 4) Assisted Human Operations, and 5) Network-Enabled, Cyber Hardened Weapons.[65] Deputy Defense Secretary Robert Work, at a speech in 2016, made it clear that the core "…technological sauce of the Third Offset is going to be advances in Artificial Intelligence (AI) and autonomy."[66]

Signaling DoD's commitment to the Third Offset Strategy and the focus areas of the DoD AI Strategy, the Joint Artificial Intelligence Center was created in 2018. The Joint Artificial Intelligence Center was created to be a "focal point of the DoD AI Strategy," and was established to "accelerate the delivery of AI-enabled capabilities, scale the Department-wide impact of AI, and synchronize DoD AI activities to expand Joint Force advantages."[67] Mimicking its innovative support to previous offset strategies, the Defense Advanced Research Projects Agency (DARPA) is supporting the current offset with several AI initiatives that include "streamlined contracting procedures"[68] designed to entice AI researchers into accelerated contracts for rapid AI research and innovation – for which they've announced a $2 billion dollar funding stream.[69]

Most pertinent to the premise of this paper is the work DARPA is undertaking to figure

---

[65] Katie Lange, "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are," *DoDLive*, 30 March 2016, http://www.dodlive.mil/2016/03/30/3rd-offset-strategy-101-what-it-is-what-the-tech-focuses-are/.
[66] Robert Work, "Remarks by Deputy Secretary Work on Third Offset Strategy: Delivered Brussels, Belgium," Department of Defense, last modified April 28, 2016, https://dod.defense.gov/News/Speeches/Speech-View/Article/753482/remarks-by-d%20eputy-secretary-work-on-third-offset-strategy/.
[67] Department of Defense, *Summary of the 2018*, 9.
[68] "Accelerating the Exploration of Promising Artificial Intelligence Concepts," DARPA, last modified July 20 2018, https://www.darpa.mil/news-events/2018-07-20a.
[69]"DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies," DARPA, September 7, 2018, https://www.darpa.mil.news-events/2018-09-07.

out how to engender trust between the military operator and the machine. In the *Explainable Artificial Intelligence* initiative, researchers are seeking to "create a suite of machine learning techniques that…enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."[70] DARPA is additionally researching "whether human pilots can trust robot wingmen in a dogfight"[71] in their new Air Combat Evolution (ACE) program. The program "aims to increase warfighter trust in autonomous combat technology by using human-machine collaborative dogfighting as its initial challenge scenario."[72] Much of the innovation necessary to support the Third Offset Strategy, and the new DARPA initiatives, will require careful study and understanding of human judgment and decision-making in military contexts.

Human Judgment and Decisions Regarding the Employment of Lethal Force

In addition to the shrinking time for decision and action in warfare, there is no guarantee that having humans in the loop will always produce the best outcome. In fact, the presence of a human operator in the loop may induce error, not prevent it. A useful case involving the Aegis Weapons System demonstrates how humans-in-the-loop can make judgment errors with lethal, and strategic, consequences. "On July 3rd, 1988, USS VINCENNES [a Ticonderoga-class guided missile cruiser equipped with the Aegis Weapon System] shot down Iranian Air flight 655"[73] shortly after take-off, killing all 290 passengers aboard. The U.S. Government contended that it was an accident, a case of mistaken identity. From the perspective of what the on-scene

---

[70] David Gunning, "Explainable Artificial Intelligence," DARPA, accessed May 19, 2019, https://www.darpa.mil/program/explainable-artificial-intelligence.
[71] Patrick Tucker, "US Military Testing Whether Human Pilots Can Trust Robot Wingmen in a Dogfight," Defense One, May 7, 2019, https://www.defenseone.com/technology/2019/05/us-military-testing-whether-human-pilots-can-trust-robot-wingmen-dogfight/156817/.
[72] "Training AI to Win a Dogfight," DARPA, May 8, 2019, https://www.darpa.mil/news-events/2019-05-08.
[73] David K. Linnan, "Iran Air Flight 655 and Beyond: Free Passage, Mistaken Self-Defense, and State Responsibility," *Yale Journal of International Law* 16, no 2 (1991), 246, http://digitalcommons.law.yale.edu/yjil/vol16/iss2/2.

commander could reasonable know at the time, firing upon Flight 655 was justified as self-defense. VINCENNES and another ship, USS MONTGOMERY, were in the Persian Gulf during the end of the Iran-Iraq War, which had evolved in the maritime commons into a "Tanker War" by 1987, when the United States escorted Kuwaiti oil tankers through the Strait of Hormuz. Tensions were extraordinarily high, with threats of mines, hostile aircraft, and Iranian small boats equipped with an assortment of lethal weapons. In this 1988 incident, the U.S. was a neutral party,[74] with presence in the Arabian Gulf for the protection of merchant shipping. This scenario illuminates how humans-in-the-loop can make perceptual and inferential errors with deadly results, particularly in ambiguous, gray zone operations.

The morning of 03 July, 1988, VINCENNES and MONTGOMERY were responding to an incident of alleged hostile fire from an Iranian Revolutionary Guard (IRG) speedboat on a helicopter from VINCENNES. Not long after arriving on scene and operating in Iranian territorial waters, VINCENNES began firing on IRG speedboats. Simultaneously, Iranian Air Flight 655 departed late from Bandar Abbas air base en-route to Qatar. It was assigned a Mode III IFF frequency denoting it as a civilian airliner. Flight 655 climbed consistently toward its assigned flight altitude of 14,000 ft, and kept strictly to an assigned international, commercial aviation corridor. VINCENNES' Aegis AN/SPY-1 radar detected the commercial aircraft upon take-off, and in the stress of the engagement with IRG speedboats, VINCENNES Combat Information Center (CIC) personnel mis-attributed a Mode II military IFF frequency to the Flight 655 track. VINCENNES tried to hail the aircraft on International and Military Air Distress channels to confirm its identity. However, when VINCENNES received no reply, the personnel in CIC classified the track as a hostile Iranian F-14. The VINCENNES' Commanding Officer

[74] Linnan, "Iran Air Flight 655," 268.

(CO) relied upon information regarding the contact as provided by his Tactical Action Officer and CIC, and enmeshed it in the context of a hostile situation where he was engaged in a gun battle with IRG speedboats. In this context, the humans-*in*-the-loop were unable to accurately distinguish between a civilian airliner, exhibiting the appropriate flight profile and identification frequencies, and a hostile Iranian military aircraft; "…CIC personnel responsible for air defense misinterpreted significant portions of the objective data."[75] In contrast, USS SIDES – a guided missile frigate – which was operating 18 nautical miles away from VINCENNES, and in closer proximity to the Bandar Abbas airbase, correctly identified the contact as a civilian commercial flight. USS SIDES operated a less advanced, long-range air-search radar, the AN/SPS-49. Despite the less advanced sensor system, the SIDES CO, CDR David Carlson, decided that the air contact was not a threat, stating in a 2000 BBC documentary interview that "it did not meet any of the threat parameters."[76]

After receiving no response to repeated hails via distress channels, the VINCENNES fired upon Flight 655. In documentary video filmed on the day of the incident, and later incorporated into a BBC documentary, VINCENNES personnel can be seen and heard cheering on the bridge of the ship after its missiles impact the commercial aircraft.[77] In its argument before the International Court of Justice, the U.S. claimed that VINCENNES was exercising its right of self-defense and, though tragic, the downing of Flight 655 was "incident to the lawful use of force." However, David Linnan in the 1991 Yale Law Journal review of downing of Flight 655 by an Aegis launched missile calls it a case of "mistaken self-defense," which did "not excuse the use of force." This is a critical perspective because it could provide some insight

[75] Linnan, "Iran Air Flight," 252.
[76] "US Missile shoot down - Iran Air Flight 655 Documentary," Magnetpraetorian, video, 41;45, 2000, accessed May 25, 2019, https://www.youtube.com/watch?v=lRJnumxuHwY.
[77] "US Missile shoot down - Iran Air Flight 655 Documentary."

that *even with* humans-*in*-the-loop, and even in only self-defense cases, the opportunity for devastating error is possible. When investigators reviewed the computer records from USS VINCENNES after the incident, they found that the system had correctly classified the track based on IFF frequency, and contrary to CIC reporting, held the track to be constantly climbing in altitude. Had the humans relied on the machine in this situation, 290 people might have arrived at their destination unharmed. In the same BBC documentary, the SIDES CO commented on the possibility of being overwhelmed by information:

> *You were inundated with intelligence messages, projecting the worst-case scenario possible, every day of the week. Such that, if anything happened, whatsoever, they could go back to the file and pull out a warning that said that, 'well, we've warned them about that.' But, life in the Gulf was business-as-usual. Commerce continued, airliners continued to fly back and forth. If you allowed yourself to focus solely on those intelligence reports, without going up on deck, walking around and looking at the reality of life in the Persian Gulf, you could become quite paranoid about threats that didn't exist.*[78]

As this case demonstrates, the *appropriate* level of human judgment may be difficult to define, will be variable given the operating environment, and will be vastly more constrained by the factor of time. In the investigative report on this incident, the Investigative Officer, Admiral William Fogarty, remarked on the compressed reaction time as a factor affecting the decision making of VINCENNES CO:

> *Time compression played a significant role in the incident. From the time the CO first became aware of TN 4131 [Iranian Air Flight 655] as a possible threat, until he made his decision to engage, the elapsed time was approximately three minutes, 40 seconds. Additionally, the Commanding Officer's attention which was devoted to the ongoing surface engagement against IRGC forces (the "wolf closest to the sled"), left very little time for him to personally verify information provided to him by his CIC team- a team in which he had great confidence. The fog of war and those human elements which affect each individual differently-not*

[78] "US Missile shoot down - Iran Air Flight 655 Documentary"

> *the least of which was the thought of the Stark incident—are factors that must be considered.*[79]

The VINCENNES case is ideally suited for additional analysis, even thirty years on, as the DoD progresses with integrating AI and autonomous technology into the force. An early example of gray zone operations, the investigation of this case provides evidence that suggests that trusting in the machine would have reduced ambiguity and improved command decision-making. Admiral Fogarty remarked that, "The AEGIS Combat System's performance was excellent - it functioned as designed. Had the CO USS VINCENNES used the information generated by his C&D system as the sole source of his tactical information, the CO might not have engaged TN 4131 [Iranian Air Flight 655]."[80]

In contrast to ill-defined levels of human judgment required to employ AI and autonomous technology, an aspect of strategy and directives governing AI and autonomous weapons which has *consistent* practical precedent is the law of war, and the expectation of military forces to adhere to laws and treaties governing warfare. The law governing the conduct of military forces in war is referred to in the military as the Law of Armed Conflict (LOAC). It is considered International Public Law, and is "also referred to as the law of war (LOW) or international humanitarian law (IHL)."[81] LOAC/IHL applies to forces that are already engaged in conflict, or *jus in bello*. The law provides "four legal principles govern modern targeting decisions: (1) Military Necessity, (2) Distinction, (3) Proportionality, and (4) Unnecessary Suffering/Humanity."[82] Some law scholars suggest that the LOAC provides "an appropriate

---

[79] William M. Fogarty, "Formal Investigation Into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988," Office of the Chairman of the Joint Chiefs of Staff, August 19, 1988, https://www.jag.navy.mil/library/investigations/VINCENNES%20INV.pdf, 78.
[80] William M. Fogarty, "Formal Investigation," 78.
[81] LCDR David Lee, JAGC, USN, ed., *Law of Armed Conflict Deskbook 2015*, 5th ed. (Charlottesville, VA: U.S Army Judge Advocate General's Legal Center and School, 2015),8 The Judge Advocate General's Legal Center and School, 2015, accessed 26 May 2019, http://www.loc.gov/rr/frd/Military_Law/pdf/LOAC-Deskbook-2015.pdf.
[82] Lee, *Law of Armed Conflict*, 133.

general framework" for "international regulation of autonomous weapon systems,"[83] rather than the extreme solutions contained in proposals that seek to ban autonomous weapons altogether. An area of particular concern for the battlefield use of AI and autonomous technology is the principle of distinction. As tragically demonstrated in the Iranian Air Flight 655 disaster, human judgment regarding the protection of non-combatants has not been flawless.

Distinction – Protection of Non-combatants

> *Incorporating AI also has the potential to enhance our implementation of the Law of War. By improving the accuracy of military assessments and enhancing mission precision, AI can reduce the risk of civilian casualties and other collateral damage.*[84]

The principle of distinction requires that military commanders, to the best of their ability, ensure that targets are military in nature – either human combatants or physical targets like buildings – and not civilian persons or property. Additional Protocol 1 of the Geneva Conventions, article 48, specifies "in order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives."[85] This principle does not suggest that there shall be no civilian casualties in conflict, but ensures that they should "never be deliberately targeted."[86] Though the United States has not ratified Additional Protocol 1 (AP1) (the U.S. is a signatory only), the customary practice of the principle of distinction in the

---

[83] Kenneth Anderson et al., "Adapting the Law of Armed Conflict to Autonomous Weapon Systems," *International Law Studies* 90, (2014): 411, https://apps.dtic.mil/dtic/tr/fulltext/u2/a613290.pdf.
[84] Department of Defense, *Summary of 2018,* 6.
[85] "Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)," 8 June 1977, *International Committee of the Red Cross*, accessed 30 May 2019, https://ihl-databases.icrc.org/ihl/WebART/470-750061?OpenDocument.
[86] Bryan Frederick and David E. Johnson, *The Continued Evolution of U.S. Law of Armed Conflict Implementation: Implications for the U.S. Military* (Santa Monica, CA: RAND, 2015): 5, https://www.rand.org/content/dam/rand/pubs/research_reports/RR1100/RR1122/RAND_RR1122.pdf.

U.S. military can be traced to General Order No. 100, issued by President Lincoln during the U.S. Civil War – also called the Lieber Code after law professor Francis Lieber who wrote the instructions.[87] In recent conflict, the U.S. has implemented operational directives and rules of engagement (ROE) that are considerably more stringent than AP1. As the former General Counsel to the Department of Defense notes, complying with LOAC in military conflict is not only required of U.S. military forces, but "complying with the law also helps us defeat our adversaries and their ideology, because it helps to confer legitimacy on our actions in the eyes of people around the world."[88]

With political and ideological pressure to conform as tightly as possible to the LOAC in increasingly complex and ambiguous conflict scenarios (as in the gray zone), and extraordinary volumes of knowable information available, U.S. military commanders must become adept at understanding probabilistic outcomes. Clausewitz says that commanders should, "be guided by the laws of probability," [89] but how well do people, and military leaders in particular, understand the laws of probability and make decisions under conditions of uncertainty?

The answer is: not very well. Amos Tversky and Daniel Kahneman, in their landmark publication, *Judgment under Uncertainty: Heuristics and Biases*, assess that people reliably use "heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations."[90] Simple judgment operations have had evolutionary survival utility for humans for many millennia, and often attain the level of intuition, but

---

[87] Francis Lieber, "General Orders No. 100: The Lieber Code: Instruction for the Government of Armies of the United States in the Field," Yale University, accessed on April 22, 2019, http://avalon.law.yale.edu/19th_century/lieber.asp, Art. 22, 23.
[88] Jennifer O'Connor, "Applying the Law of Targeting to the Modern Battlefield," Department of Defense, 2, accessed March 24, 2019, https://dod.defense.gov/Portals/1/Documents/pubs/Applying-the-Law-of-Targeting-to-the-Modern-Battlefield.pdf.
[89] Clausewitz, Howard, and Paret, *On War*, 117.
[90] Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (September 27, 1974): 1124.

"sometimes they lead to severe and systematic errors."[91] Tversky and Kahneman identified three common heuristics used in human decision-making: representativeness, availability, and anchoring.

The first heuristic or cognitive bias that leads humans astray is the representative heuristic. The representative heuristic is expressed when the probability, or likelihood of occurrence of A (e.g. an event, a description, a person, a group, etc…) is judged "by the degree to which A is representative of, or resembles B."[92] In the case of Iranian Flight 655, VINCENNES Commanding Officer and crew saw Iran as a threat. The dual military-civilian use of the Bandar Abbas airbase meant that operators could have judged it more likely that any aircraft departing from that particular airfield was a military aircraft, simply because the military and civilian aviation shared the airfield.

The second cognitive bias that systematically causes human error is the availability heuristic. The availability heuristic operates in "situations in which people assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind."[93] The hostile gunfire engagement of VINCENNES and MONTGOMERY with IRG speedboats at the time same time Flight 655's track appeared on radar made it more likely that VINCENNES crew would evaluate the track as hostile, not only because they were embroiled in a hostile engagement with the Iranians, but also because another U.S. ship, the USS STARK, had been hit and damaged by two air-launched, Iraqi Exocet missiles in the Persian Gulf only a year prior.[94]

---

[91] Tversky and Kahneman, "Judgment," 1124.
[92] Tversky and Kahneman, 1124.
[93] Tversky and Kahneman, 1127.
[94] William J. Crowe, "Formal Investigation Into the Circumstances Surrounding the Attack on the USS STARK (FFG-31) on 17 May 1987," Office of the Chairman of the Joint Chiefs of Staff, September 3, 1987, https://www.jag.navy.mil/library/investigations/USS%20STARK%20BASIC.pdf.

The third cognitive bias impairing human judgment is the anchoring heuristic. The anchoring heuristic operates when people fail to adjust their judgments or estimates of outcomes from a starting value or intuition, despite equal or greater probability the outcome is not the initial value.[95] In the case study on the Patriot Missile System, operators failed to adjust their trust and reliability in the system after the first proven incident of friendly fire, and repeated warnings from returning pilots that the fire-control radar had illuminated their aircraft. They couldn't adjust expectations from their doctrinal training, and that judgment error unfortunately cost a Navy pilot his life.

Tversky and Kahneman's studies and conclusions were limited in scope and operated in conditions of relative calm and simplicity. In the face of stressful and complex decisions involving life and death, military members may be more likely to fall back on unconscious judgment heuristics. So far, the cases reviewed have been focused on use of, or trust in, AI and autonomy. In order to elucidate how the judgment heuristics discussed above impact human decision-making in non-autonomous situations, it is appropriate to cover a case of platoon-level decision making without AI or autonomy in a combat zone. In *Redefining the Modern Military*, author H.M. Denny described a crisis situation in 2008 in which he made a decision regarding the use of force at a combat outpost in Afghanistan.[96] In this case, there were possible enemy ground forces in the vicinity of a platoon that had just struck an IED while on patrol. The possible enemy forces could not be positively identified as such, but a combat helicopter on scene requested permission to engage. Denny, a Lieutenant at the time, authorized the engagement without the delegated authority to do so according to regional rules of engagement.

---

[95] Tversky and Kahneman, "Judgment," 1128.
[96] H. M. Denny, "Professionals Know When to Break the Rules," in *Redefining the Modern Military: The Intersection of Profession and Ethics*, ed. Nathan K. Finney and Tyrell O. Mayfield (Annapolis, MD: Naval Institute Press, 2018), 53.

In the process of making this decision, he listed a series of questions that he desired answers to in order to make his decision – twelve questions he would never have a certain answer to in the time required for effective engagement of the possible enemy forces. Though the uncertainty multiplied, Denny authorized the use of force.

The outcome was successful, but Denny got lucky. The men were posthumously identified as enemy insurgents, and Denny's platoon was able to collect vital intelligence from the artifacts they were carrying. Denny states that he "was willing to accept the potential consequences, and believed I had the best situational awareness to make a decision…my professional responsibility required me to make an immediate decision that was professionally wrong."[97] Given limited information, and an inability to communicate with on-scene personnel, he made a risk decision to authorize the use of lethal force.

This example illustrates the "pervasiveness of risk and uncertainty in decision making,"[98] and holds all the elements of a difficult decision. Denny felt the decision he made to authorize engagement of potential enemy insurgents "reinforced the lessons of self-improvement" and "strengthened the decision-making processes with regard to weapons implementation."[99] The danger in this case is assessing Denny's decision as the *right* decision because the outcome was what he expected it would be. Tversky and Kahneman describe this as the "illusion of validity," and it happens when "unwarranted confidence…is produced by a good fit between the predicted outcome and the input information."[100] This case is one in which various human judgment heuristics could have contributed to the same choice with a different outcome.

---

[97] Denny, "Professionals," 56.
[98] Rose McDermott, *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy* (Ann Arbor, MI: University of Michigan Press, 1998), 5.
[99] Denny, "Professionals," 57.
[100] Tversky and Kahneman, "Judgment," 1126.

The *representativeness heuristic*[101] might have influenced Denny's assessment of the likelihood that four military-aged-males, discovered near an IED explosion, were in fact enemy insurgents, and not local, curious villagers who happened to be in the wrong place at the wrong time. His confidence in his prediction that the military-aged-males were insurgents was based on highly uncertain evidence.

The *availability heuristic*[102] would suggest that the crisis situation Denny was facing would trigger his memories of past situations and the "experience gained through numerous skirmishes, troops in contact situations, and fire missions executed,"[103] and might have influenced his assessment that the situation he was facing was just like previous encounters. He may have fallen victim to "illusory correlation" wherein his judgment was biased by his assessment of how frequently IED blasts co-occurred with visible enemy insurgents in the immediate vicinity, and the tactical "associative bond between them."[104]

The *anchoring heuristic*[105] would suggest that the initial report of the situation which led with "…lead vehicle destroyed by IED, 9-Line to follow,"[106] could have skewed Denny's perception of the relevance of the later report of possible enemy insurgents near the location of the IED explosion. The bias and heuristic influences on judgment may be outwardly expressed by military personnel as 'professional experience,' and rightfully so – decades of history reinforce the understanding that previous combat experience allows military members to perform better in subsequent combat, but the situation described by Denny had a significant probability of turning out very differently. Moreover, if the unidentified males were not insurgents, Denny

---

[101] McDermott, *Risk-Taking*, 6.
[102] McDermott, 7.
[103] Denny, "Professionals," 57.
[104] Tversky and Kahneman, "Judgment," 1128.
[105] McDermott, 7.
[106] Denny, 53.

would have violated the law of armed conflict. He stated that he was willing to accept the consequences, but in the short time span of decision to authorize an engagement Denny could not have really comprehended the possibility of error and the consequences of that error.

And what if the military commander has had no previous experience? What if the prediction of future naval combat implores us to recognize that a warship will likely be overcome by a saturation attack of missile salvos, aircraft strikes, and torpedoes, but there is no way to test a commander's response, nor allow him or her to work through and become self-aware of all the judgment biases that may cloud his or her assessment of the combat situation?

Judgment errors and cognitive biases plague human, military decision making regardless of whether the engagement is by a human or a machine. The exploration of the cases involving the military's use of the Aegis Weapons System and the Patriot Missile System suggest that, although the U.S. military has been actively employing narrow artificial intelligence in lethal, semi-autonomous weapons systems, the employment of the weapons system through a full detect-to-engage sequence has been limited to situations of self-defense. When asked under what conditions a ship Commander would authorize supervised autonomy for the Aegis system, interviewed respondents suggested it would only be allowed if the ship and crew were in mortal danger. Sacrificing command authority to an intelligent system that can make targeting and engagement decisions faster than a human seems like a logical step when facing an incoming salvo of supersonic missiles, but what judgment biases and heuristics will impair the Commander's ability to assess the existence of a true existential threat? Given the right context, everything may look like a threat, or nothing at all. If the existential threat response requires a moral pause[107] where survival supersedes distinction and the LOAC, the triggers for this

[107] Hans Jonas, "Toward a Philosophy of Technology," *The Hastings Center Report* 9, no 1 (New York: Hastings Center, 1979): 24.

response must be contemplated, understood, and exercised well before commanders are faced with making that choice.

The pattern becoming clear throughout this research is a consistent default to humans-*in*-the-loop across the spectrum of conflict, though humans are prone to judgment errors. The consist barrier to enabling AI and autonomy in a military context is a lack of trust. The use of force in Afghanistan, covered in the preceding paragraphs, provide an example of human judgment errors within the context of international armed conflict where LOAC clearly applies. Yet even in that context, distinction between civilians and combatants is difficult. In peacetime, it is clear that policy, doctrine, and operator preference guide the force to trust the human. In war (as in the Patriot case), the normative response is to trust the machine – except when the machine performs in an unpredictable or unreliable manner, in which case the force should trust the human. What about the gray zone? In the gray zone, the force should be encouraged to trust the machine because the compressed reaction times, and likelihood of escalation may preclude effective self-defense. But "it is not clear when gray zone conflicts stop being conflicts at all and start becoming something else, something that we don't yet understand or have words to describe."[108] To further compound the problem, the risk of escalatory action based on a machine-induced accident, coupled with the inability to distinguish between civilians and combatants, requires ROE that directs the force to trust the human, not the machine.

The United Nations (UN) has attempted to improve global governance of autonomous weapons development in accordance with the LOAC. Compliance with the principle of distinction in the LOAC is not only required by law, but also provides additional effectiveness in

[108] Nora Bensahel, "Darker Shades of Gray: Why Gray Zone Conflicts Will Become More Frequent and Complex," Foreign Policy Research Institute E-Notes, February 13, 2017, https://www.fpri.org/article/2017/02/darker-shades-gray-gray-zone-conflicts-will-become-frequent-complex/.

the achievement of military objectives as the DoD General Counsel stated. Members of the UN Convention on Certain Conventional Weapons (CCW) attempted to begin work on a treaty for the ban of fully-autonomous weapons, but the resolution was blocked by the U.S., Russia, South Korea, Israel, and Australia.[109] The DoD AI Strategy emphases that the utilization of systems, autonomous or intelligent, that can improve a military decision-maker's judgment in combat, and accuracy in distinction of military targets should be explored. The U.S. expressed this sentiment in addressing the United Nations 2018 Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS). The U.S. outlined its understanding that autonomy in weapons systems should ensure that commander's (human's) intentions should be carried out, with emphasis on the ability of "personnel to exercise appropriate levels of human judgment over the use of force."[110] The submission goes on to cite examples currently deployed in the force (like the Aegis Weapons System and Patriot Missile System) where autonomous targeting functions are more appropriate – in fact preferred – over manual, human control due to speed and accuracy in targeting and engagement. The submissions predict that, "as technology advances […] autonomous weapons will enjoy greater capability to comply with legal obligations, and, in some situations, may out-perform humans in this regard."[111] This response sets up a condition where the U.S. recognizes the capabilities of the Aegis and Patriot Systems exceed human performance, but suggest that those capabilities will necessarily be required to be limited by an

---

[109] Mattha Bussby and Anthony Cuthbertson, "Killer Robots Ban Blocked by US and Russia at UN Meeting," *The Independent*, September 3, 2018, https://www.independent.co.uk/life-style/gadgets-and-tech/news/killer-robots-un-meeting-autonomous-weapons-systems-campaigners-dismayed-a8519511.html.

[110] Group of Governmental Experts on Lethal Autonomous Weapons Systems, "Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," August 28, 2018, 1, https://www.unog.ch/80256EDD006B8954/(httpAssets)/D1A2BA4B7B71D29FC12582F6004386EF/%24file/2018_GGE+LAWS_August_Working+Paper_US.pdf.

[111] Dan Saxon, "A Human Touch: Autonomous Weapons, Directive 3000.09, and the 'Appropriate Levels of Human Judgement over the Use of Force,'" *Georgetown Journal of International Affairs* 15, no. 2 (Summer/Fall 2014):102.

organic timescale and judgment biases of human commanders, who bear the responsibility of ensuring the weapons systems are employed "…with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable rules of engagement (ROE)."[112] This implies human accountability for *all* AI or autonomous weapons violations of the law of war, treaties, safety rules, and ROE, intended or not. What does the human require from the machine, or machine designers, to accept this level of accountability? Trust.

Trust as a Barrier to Implementing AI and Autonomy in Warfare

Trust as a Function of Predictability

A strong component of trustworthiness is predictability, which is often measured by determining how often the outcome of a decision or action achieved the expected results. Some researchers suggest that there are two criteria of trust, reliance and perfect confidence.[113] Reliance implies a certain knowledge of agent A's ability to perform a certain action, X.[114] Reliance (or reliability) and confidence come together to make up the concept of predictability. In the case of human-machine interaction, it is desirable to review this aspect of trust for both a machine and a human. Predictability was foremost on the minds of engineers during the mid-1980's development of Naval combat systems (Aegis) that operated on doctrine statements. Pursuit of predictability induced the Naval Sea Systems Command to create a Doctrine Working Group, which outlined several foundational reasons for standardization of "doctrinal automation." Principal among them was a requirement for "predictable, desired response."[115] In

[112] Department of Defense, *Autonomy in Weapon Systems*, 3.
[113] H. J. N. Horsburgh, "The Ethics of Trust," *The Philosophical Quarterly* 10, no. 41 (Oct 1960): 344.
[114] Horsburgh, "The Ethics of Trust," 344.
[115] Gersh, "Doctrinal," 76.

essence, this meant that the combat system's response to a given "tactical situation must be predictable to ship, warfare area command, and composite warfare command personnel."[116]

In the case of Aegis doctrine statements, this level of predictability seems inherent in the transparency of the if/then statements written by ship Commanders, and incorporated in the Command and Decision (C&D) element of the Aegis Weapons System. However, this simplicity may belie a level of complexity in the Aegis system that hasn't been questioned because operators have not had first-hand experience with such complexity since the first Aegis ship rolled off the docks. Aegis, as a defensive weapon system, is ostensibly designed for responding, fully autonomously, to an incoming missile attack, particularly one with multiple tracks in a coordinated naval salvo. What many Commanders likely don't fully comprehend is how Aegis' detection, classification, targeting, and engagement system prioritizes and inter-relates responses to multiple inbound tracks.

> A set of doctrine statements can interact with each other in complicated ways, since the action of one statement, like identification, can be a criterion used by another statement, like one controlling engagement. In addition, the details of a combat decision system's internal processing of doctrine statements (the exact ways in which track parameters are compared with doctrine statement criteria, the timing of the comparisons, and the internal logic used to resolve conflicts and set evaluation priorities) can at times produce unexpected results.[117]

In an interview with Bradford Tousley, director of DARPA's Tactical Technology Office, author Paul Scharre discusses the director's primary concern with fielding autonomous systems which is the ability to demonstrate system reliability through test and evaluation: "What I worry about the most is our ability to effectively test these systems to the point that we can quantify that we trust them. Unless the combatant commander feels that the autonomous system

[116] Gersh, "Doctrinal," 76.
[117] Gersh, 76.

is going to execute the mission with the trust that he or she expects, they'll never deploy it in the first place."[118]

Trust as a Function of Knowledge and Transparency

Humans have a hard time trusting things they don't understand. Developing trust in AI and autonomous systems means "interacting with something we don't understand [which] can cause anxiety and make us feel like we're losing control."[119] Even in the early adoption of the Aegis Weapons System, engineers and tacticians were concerned with the ability of a human operator to fully understand the complexity of the actions the system performed. In a response to Dr. Gersh's article, a rather prescient assessment by Michael Lindemann from the Naval Surface Weapons Center compares human ability to the AWS. Lindemann observes that "complexity inhibits understanding."[120] Noting that the AWS doctrine statements could number as many as seventy-five, he remarked that "there is no simple way for the commander, or operator, to gain a comprehensive understanding of the active set of doctrine statements, and thus, a clear comprehension of its potential resultant action."[121]

Julia Macdonald and Jaquelyn Schneider presented interesting findings on human-machine trust from their surveying over 400 Joint Tactical Air Controllers (JTACs) and Joint Fires Observers (JFOs) regarding their perception of trust in unmanned drones performing close air support (CAS). Those surveyed overwhelmingly preferred manned aircraft performing CAS. In their findings, they exposed the difference between human confidence in the machine's ability

---

[118] Bradford Tousley, quoted in Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, 2018), 83.
[119] Vyacheslov Polonski, "People Don't Trust AI--Here's How We Can Change That," *Scientific American*, January 10, 2018, https://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/?redirect=1.
[120] Michael J. Lindemann, response to "Doctrinal Automation in Naval Combat Systems: The Experience and The Future," *Naval Engineers Journal* 99, no. 4 (July 1987): 108.
[121] Lindemann, response to "Doctrinal Automation," 108.

to effectively perform as designed and reliability of the machine for use in a designed mission. The question they were unable to answer – and one which continues to be key – is "at what point confidence in these machines becomes high enough to create trust?"[122]

Much of the AI employed in the DoD is narrow AI with deterministic logic. In a deterministic model, the output is determined solely by the initial conditions and the values of the parameters being modeled. There is no random variability, and mathematical traceability is generally possible. The early forms of automation and machine intelligence consisted of an indeterminate number of if-then statements, which are generally considered deterministic. In a deterministic environment, "the next state of the environment is completely determined by the current state and the action executed by the agent."[123] In the case of AI, 'if' the software or machine encounters a specific range or set of parameters or variables, 'then' it will perform an action to effect an expected outcome. These if/then actions are programmed linearly using finite algorithms and often attempt to mimic a simple human logical assessment. However, the real world, and especially war, is stochastic – highly uncertain and unpredictable. Future AI, even narrowly scoped, will be take goal-oriented action with much less traceability of decision paths (especially with deep learning and neural networks). Future AI will be able to learn independent of human input, and may have stochastic responses to complex wartime context. "It is becoming increasingly clear that human beings may not necessarily always be able to understand how (and

---

[122] Julia Macdonald and Jacquelyn Schneider, "Trust, Confidence, and the Future of Warfare," *War on the Rocks*, February 5, 2018, https://warontherocks.com/2018/02/trust-confidence-future-warfare/. The authors have faced significant scrutiny of their findings, most ardently from a cadre of former and current MQ-1 and MQ-4 drone pilots. While their survey results likely contain artifacts from the particular slice of time they surveyed (and they admit as much), I found the rebuttal by Cory T. Anderson et al., in "Trust, Troops, and Reapers: Getting 'Drone' Research Right," (War on the Rocks, April 3, 2018, https://warontherocks.com/2018/04/trust-troops-and-reapers-getting-drone-research-right/) to have significant inconsistencies in reference usage, and limited to erroneous application of 'refuting' data. What Macdonald and Schneider are seeking to illuminate (trust) may not be able to be established by data on multi-mission platform, especially when the primary mission for drone development was ISR.

[123] Russell and Norvig, *Artificial*, 43.

possibly why) autonomous systems make decisions,"[124] which may increase human distrust of machine decision making, and decrease the ability for assigning accountability for mistakes.

Trust as a Function of Accountability, Morality and Ethics

"*Whoever sheds the blood of man, by man shall his blood be shed...*"[125] – *Genesis 9:6*

Philosophers and theologians understand the decision to deliberately take a human life as a moral decision. The sanctity of life is a consistent theme in the teachings of the world's three largest religions, Christianity, Islam, and Hinduism. For those of faith, the morality of killing – the rightness or wrongness of the action – seems to be a concept easily accessible. Even for those who are agnostic or atheist, the morality or 'wrongness' of killing is grasped by intuition and reflected in norms of reciprocity.

For millennia, men and women have wrestled with the implications of the moral decision to take human life in warfare.[126] In the religious texts for the above-mentioned faiths, a critical component that accompanies the sanctity of life is the accountability and punishment to be assigned when the maxim is deliberately violated. The arguments and doctrine presented in religious teachings and philosophical work create a body of knowledge commonly referred to as Just War Theory, which often recalls the work of St. Thomas Aquinas as an inflection point in history for all Law of Warfare theories that follow.[127] The introduction in recent decades of the possibility that non-human entities could make what we consider implicitly to be a human moral decision is a situation for which we don't have a body of work to guide our path, but is also a situation that the DoD, through directives and policy, does not intend to allow.

---

[124] Kenneth Anderson et al., "Adapting the Law of Armed Conflict to Autonomous Weapon Systems," *International Law Studies* 90, (2014): 394, https://apps.dtic.mil/dtic/tr/fulltext/u2/a613290.pdf.
[125] The New Student Bible, New International Version (Grand Rapids, MI: Zondervan Publishing House, 1986), Genesis, 9:6, 33.
[126] Michael Walzer, *Just and Unjust Wars* (New York: Perseus Books Group, 1977), 3-20.
[127] William H. Shaw, *Utilitarianism and the Ethics of War* (New York: Routledge, 2016), 14.

As DoD works to realize its 2018 AI strategy, it must develop a keen understanding of factors of human behavior that influence human-machine interaction. Recent research into human perceptions of morality, trustworthiness, and accountability highlight a conundrum that will continue to vex developers of AI and autonomous systems: humans tend to hold machines to different moral standards of behavior than they do other humans. Experiments have shown that humans intuitively and socially prefer other humans who exhibit deontological morality in decision making than those who exhibit utilitarian morality. For example, someone who believes that stealing is always wrong so they never steal (deontological approach) may be more trustworthy than someone who believes that stealing may be okay if the consequences, or outcome, is of great benefit to the greatest number of persons (utilitarian approach).[128] In this research, humans who exhibited deontological behavior in their deliberate actions were deemed more trustworthy. The researchers used two versions of the classic trolley car dilemma: the trolley car, or 'switch' experiment, and the 'footbridge' experiment. In both cases, the experiment subject is faced with a choice to either authorize the death of one person to save five, or do nothing and allow five people to die when a trolley car crashes. The difference between the two is critical. In the switch experiment, the subject need only throw a mechanical switch and the trolley car changes from running on a track that will kill five people, to running on a track that will kill one person. In the footbridge experiment, however, the choice between one and five deaths is modified, and the subject must push one person off a footbridge to certain death in the path of a runaway trolley with five people on board.

The experiments found that "participants perceived the deontological agent [kill one person] to be more trustworthy in the footbridge dilemma, but not the switch dilemma," and that

---

[128] Jim A.C. Everett et al., "Inference of Trustworthiness from Intuitive Moral Judgments," *Journal of Experimental Psychology: General* 145, no. 6 (2016): 773.

"participants trusted the deontological agent more than the consequentialist agent in the footbridge dilemma but not the switch dilemma." Researchers speculate that the difference in perceived trustworthiness between the switch and footbridge dilemmas has to do with the difference in the actions required in each dilemma. In the footbridge dilemma, the subject must directly take physical action to end one person's life (push another person off a bridge onto train tracks below) in order to save the lives of the five people on the train. In this dilemma, someone who chose the deontological option [kill one person] was perceived as having used another person's life as a means to an end. What is not explored is how the removal of the direct, physical action of killing, by using a mechanical switch as in the switch dilemma, may have adjusted the perceived trustworthiness and morality of the subject.

Additional research in this area has reaffirmed the earlier thesis that humans have different expectations of accountability, and associated blame, for robots and/or autonomous machines. Though the following research really focuses on the aspect of blame, as a concept blame implies an associated expectation of moral judgment and an accountability for that judgment if it is perceived to be in error. Accountability and expectations of moral judgments are important components of trustworthiness. Researchers at Brown and Tufts University presented their work on understanding people's moral judgments of robot agents at the 2015 International Conference on Human-Robotic Interaction. When placed in "an identical moral dilemma,"[129] they found that humans expected robots [think AI and autonomous machines] to act in a manner that would sacrifice one life for the good of many lives, and "they were blamed more than their human counterparts when they did not make that choice."[130] This is opposite to

[129] Bertram F. Malle et al., "Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents," in *Proceedings of the Tenth Annual AMC/IEEE International Conference on Human-Robot Interaction* (March 2015): 117, accessed May 11 2019, https://hrilab.tufts.edu/publications/malletal15hri.pdf.
[130] Malle et al., "Sacrifice One for the Good," 117.

the higher social preference and trustworthiness that humans hold for other humans who make a deontological choice to save one life, even though it may mean others will die.

Considering both the Patriot incidents in 2003 and the Aegis Iranian Air Flight 655 incident in 1988, it is interesting to note that although humans were either *on*-the-loop or *in*-the-loop in both cases, in neither case was an individual human operator or decision maker held accountable for the machine-assisted mistakes that resulted in fratricide (2003) and the death of 290 civilians (1988). The Commanding Officer of the VINCENNES completed the ship's scheduled deployment, returned to homeport, and received a Meritorious Service Medal for his service on VINCENNES.[131] The Lieutenant in charge of the Patriot system was cleared after the Army's investigation, with the assessment that "she made the best call with the information she had."[132] Expectations of moral decision-making while employing AI and autonomous technology without accountability creates an opportunity for moral hazard, wherein military members may actually have *less* incentive to mitigate the risk of AI or autonomous weapon employment when there is a perception that they are shielded from the consequences of the decision.

Modern ethicists suggest that machines that will excel at compliance ethics,[133] and compliance ethics in the military often take on a deontological nature – absolute rules, or imperatives like 'do not kill.' This perspective may disagree in reality with the results of the previously reviewed experiments at Oxford and Tufts, which show that humans intuitively expect machines to make a utilitarian choice, but expect humans to make a deontological choice. Reason suggests that there is a distinction between ethical/legal compliance and human judgment

[131] "US Missile shoot down - Iran Air Flight 655 Documentary."
[132] Scharre, *Army of None*, 141.
[133] George Lucas, *Military Ethics: What Everyone Needs to Know* (New York: Oxford University Press, 2016), 183.

regarding morality. In this logic, special moral relationships (like trust) with machines may not be possible if machines are considered amoral.[134] In this case, humans – as moral arbiters – will never be completely replaced in moral decision making because "there are inherent moral limitations on special activities that must be attained through a kind of practical reasoning that, unlike strict legal compliance, is fuzzy and ambiguous, and can't really be programmed reliably."[135] This distinction between human morality and machine compliance may be correct, but the exclusion of a moral relationship between the two is erroneous. Machines may be amoral, and given goals, or sets of ethical rules (like the LOAC) machines will likely exceed human performance in being "safe, reliable, and legally compliant."[136] But regardless of the moral status of the machine, research and experiments suggest that there can be a trust relationship between human and machine.

Opportunities for Establishing Trust of AI and Autonomy in Warfare

The Rendulic Rule – AI-Enhanced Compliance with LOAC

Presupposing that the teleological goal of a military force, in international armed conflict, is to defeat an adversary's opposing military force, and that achieving this goal may require the use of force to cause harm, even death, to the adversary's military, the normative goal constraints[137] applied by the LOAC in order to protect civilian populations guide the appropriateness of the means used to achieve this goal. There are suggestions that humans may be cognitively unable to optimize choices regarding means and methods in decisions on the use of force against adversaries. Either humans "cannot consider all possible strategies for achieving

---

[134] Lucas, *Military Ethics,* 183.
[135] Lucas, 182.
[136] Lucas, 183.
[137] Giovanni Sartor, "Doing Justice to Rights and Values: Teleological Reasoning and Proportionality," *Artificial Intelligence Law* 18, (2010): 176.

certain objectives,"[138] or they are unable to "assess the rationality"[139] of the choices, as highlighted in the previous discussion applying Tversky and Kahneman's heuristics analysis to Mr. Denny's experience in Afghanistan. Decisions on the use of force with due consideration of proportionality and distinction require decision makers to compare "anticipated military advantages with anticipated civilian losses."[140] Given the excessive amount of information available in today's battlespace this is a difficult task for a human decision-maker to master. Like the Lieutenant in charge of the Patriot battery in OIF, military decisions-makers often make the best call with the information they have – unaware of, or unable to account for, the critical information they don't have but especially need.

In assessing proposed military action for compliance with the LOAC, particularly the principles of distinction and proportionality, commanders are expected to make decisions based on "the circumstances known to the military commander at the time after taking all feasible measures to ascertain those circumstances."[141] The Rendulic Rule "sets out the obligations of the reasonable military commander"[142] to take all feasible measures to establish an accurate assessment of the environment, but limits liability "based on the information reasonably available at the time of the commander's decision."[143]

The proliferation of sensors and associated data in the operating environment require a new understanding of what constitutes 'reasonably available,' and what extent of pursuit of information satisfies 'feasible measures.' As described by the Commanding Officer of USS SIDES when recalling the hectic operational tempo in the Arabian Gulf in 1988, a commander

---

[138] Sartor, "Doing Justice," 182.
[139] Sartor, 183.
[140] Ben Clark, "Proportionality in Armed Conflicts: A Principle in Need of Clarification?," *Journal of International Humanitarian Legal Studies* 3, no. 1 (2012): 77-78.
[141] Clark, "Proportionality," 78, fn 19.
[142] Clark, 78, fn 19.
[143] Lee, *Law of Armed Conflict*, 135.

could be inundated with intelligence, and assess threats where threats did not exist. His statement indicated that the intelligence community engaged in a covering maneuver by providing excessive quantities of information with the expectation that the Commander would be able to sort the proverbial wheat from the chaff.

The situation has only gotten worse in the ensuing decades. U.S. military forces are increasingly exposed to complex risk that may be misunderstood and lethally miscalculated. The glut of information impairs analyst's ability to create actionable intelligence. "Military drone operators amass untold amounts of data that never is fully analyzed because it is simply too much."[144] In a 2017 keynote address, the Chief of Naval Operations (CNO), Admiral John Richardson, referred to John Boyd's class Observe-Orient-Decide-Act (OODA)[145] loop when describing where the competition for advantage exists:

> *I would argue that […] because of advances in space and other areas […] the era of competition for precision is moving to an era of competition for decision superiority. And so, if you think of just the OODA loop – observe, orient, decide, act – we have really concentrated on is that first O, the…observe, right? And so, if we had better information…we could get more precision, and that would lead to better orientation, decisions and actions. But as these satellites and other sensors proliferate and become ubiquitous, …the playing field on that observe part of that cycle is really leveling out. In fact, data is becoming – you know, it's just coming in avalanches. And so it shifts the competition now to who can sift through that data, orient themselves better, and then made a decision. If everyone can observe, and the data is…just in monstrous amounts, …the quickest to figure out what matters and to make a decision is going to be the winner.*[146]

The DoD AI Strategy seeks to utilize AI to improve performance in observation and orientation. By utilizing the computational capabilities of AI, decision-making will be made

---

[144] Sandra I. Erwin, "Too Much Information, Not Enough Intelligence," *National Defense Magazine*, May 1 2012, http://www.nationaldefensemagazine.org/articles/2012/5/1/2012may-too-much-information-not-enough-intelligence.

[145] William S. Angerman, Capt., USAF, "Coming Full Circle with Boyd's OODA Loop Ideas: An Analysis of Innovation Diffusion and Evolution," (MA Thesis, Air Force Institute of Technology, March 2004), 3-4, https://apps.dtic.mil/dtic/tr/fulltext/u2/a425228.pdf.

[146] John Richardson, "Countering Coercion in Maritime Asia," (Remarks, Washington, DC: Center for Strategic and International Studies, 2017), https://www.csis.org/analysis/remarks-cno-adm-richardson.

more efficient and effective. For example, "perception tasks such as imagery analysis can extract useful information from raw data and equip leaders with increased situational awareness. AI can generate and help commanders explore new options so that they can select courses of action that best achieve mission outcomes, minimizing risks to both deployed forces and civilians."[147] Figure 2 below depicts the U.S. military's Joint Dynamic Targeting Cycle. Though generally applied to unplanned targets or "targets of opportunity,"[148] the process is applied to all offensive targeting decisions.

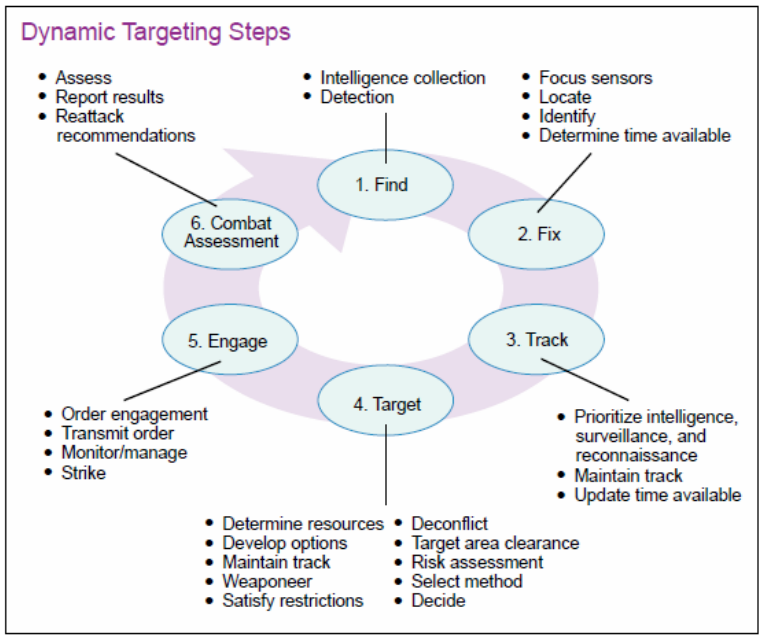

Figure 2. Dynamic Targeting Cycle. Included in Joint Publication 3-60, Joint Targeting, as Figure II-10, 28 September 2018.[149]

Operational employment cases reviewed in this research were examples where AI and autonomy were used in a defensive manner, but the same F2T2EA process (Find, Fix, Track, Target, Engage, Assess) was followed, with various levels of human intervention in all steps, but

[147] Department of Defense, *Summary of the 2018*, 11.
[148] Joint Chiefs of Staff, *Joint Targeting*, Joint Publication (JP) 3-60 (Washington, DC: CJCS, 28 September 2018), II-23.
[149] Joint Chiefs of Staff, *Joint Targeting*, II-23.

critically in step five – engagement. Trust in machines at this critical step can be improved by safely incorporating AI and autonomy in steps one through four. "Weapon systems with greater and greater levels of automation could – at least in some battlefield contexts – reduce misidentification of military targets, better detect or calculate possible collateral damage, or allow for using a smaller quanta of force compared to human decision making."[150] This becomes a pressing necessity in gray zone operations where steps one through three in the dynamic targeting cycle, and the functions of observe and orient in the OODA loop, are saturated with information and various levels of military and political posturing and signaling.

Improve Trust in the Organization

The one aspect of operationally employing AI and autonomy in the DoD that has yet to be discussed is the trust required between the DoD and the military forces supervising, assisted-by, or teamed with this new technology. Peer competitor investment in, and use of, AI and autonomous technology poses a potential threat to the security of the United States and its allies. In this competitive environment, the DoD mustn't let a race to be first in AI and autonomy, under the auspices of a Third Offset Strategy, tear the fabric of professional accountability for the safety and welfare of the force.

Wing Commander Jo Brick, an Officer in the Royal Australian Air Force, describes a special relationship of trust between the military and the state, which she terms a "fiduciary relationship."[151] A fiduciary relationship is "a relationship in which one party places special trust, confidence, and reliance in and is influenced by another who has a fiduciary duty to act for

[150] Kenneth Anderson et al., "Adapting the Law," 394.
[151] Jo Brick, "The Military Profession: Law, Ethics, and the Profession of Arms," *Redefining the Modern Military: The Intersection of Profession and Ethics*, ed. Nathan K. Finney and Tyrell O. Mayfield (Annapolis, MD: Naval Institute Press, 2018), 23.

the benefit of the party."[152] In the assessment of this special relationship, Brick outlines the state's expectations of the military, which include advice to the state on the most advantageous use of force, and strict adherence to standards of conduct.[153] However, this assessment fails to consider that any relationship succeeds or falters relative to the amount of cooperation involved. The state should only expect to be able to trust the military with fiduciary duties so long as the military can trust the state to provide "guidance to direct and constrain the disciplined application of violence for a political end."[154] This has never been more urgent a duty than now. The U.S. must ensure military leaders clearly understand when and how they may be held accountable for the risk decisions required for the employment of an artificially intelligent and autonomously capable force.

Conclusion

The development and employment of artificial intelligence (AI) and autonomous weapons in the U.S. military has progressed with the protective, doctrinal insistence that military leaders who employ such technology exercise "appropriate levels of human judgment over the use of force." However, the time available to exercise human judgment and to bring lethal force to bear has decreased while the amount of contextual information that enables decisions on the use of force has increased. At the same time, gray zone conflict activity is increasingly blurring the line between peacetime operations and warfare. U.S. military forces exerting forward, deterrent presence in areas prone to activity that is not in accordance with international norms or law are increasingly exposed to complex risk that may be misunderstood and lethally

---

[152] "Fiduciary Relationship," Merriam Webster Online Dictionary, accessed 26 May 2019, https://www.merriam-webster.com/legal/fiduciary%20relationship

[153] Brick, "The Military Profession," 27.

[154] Rebecca Johnson, "Ethical Requirements of the Profession: Obligations of the Profession, the Professional, and the Client," *Redefining the Modern Military: The Intersection of Profession and Ethics*, ed. Nathan K. Finney and Tyrell O. Mayfield (Annapolis, MD: Naval Institute Press, 2018), 95.

miscalculated. Despite strategies, directives, and invectives that eschew the need for artificial intelligence and autonomy in today's and tomorrow's battlespace, the consistent norm in the force is to default to humans *in* the loop, across the spectrum of conflict, slowing the reaction time and decision space to organic, human speeds.

Military forces must balance the duty to abide by the principle of distinction in the law of armed conflict (LOAC) and the inherent right to self-defense. AI and autonomous weapons have the potential to improve both the success of self-defensive actions and the adherence to LOAC – particularly in compressed timescales – but only if humans and organizations are able to establish trust in the machine. Establishing trust requires predictability, knowledge, and transparency of machine decisions, and clear lines of accountability for moral decisions. Addressing these considerations in the development of new AI and autonomous systems for military use will be necessary to ensure that servicemember and societal trust in the Department of Defense (DoD) is preserved, and military forces retain their will and ability to exercise lethal force.

Bibliography

Allison, Graham. “The Thucydides Trap: Are the U.S. and China Headed for War?” *The Atlantic*. September 24, 2015. https://www.theatlantic.com/international/archive/2015/09/united-states-china-war-thucydides-trap/406756/.

Anderson, Kenneth, Daniel Reisner, and Matthew Waxman. “Adapting the Law of Armed Conflict to Autonomous Weapon Systems.” *International Law Studies* 90, (2014): 386-411. https://apps.dtic.mil/dtic/tr/fulltext/u2/a613290.pdf.

Angerman, Capt. William S., USAF. “Coming Full Circle with Boyd’s OODA Loop Ideas: An Analysis of Innovation Diffusion and Evolution,” Master’s Thesis, Air Force Institute of Technology, March 2004. https://apps.dtic.mil/dtic/tr/fulltext/u2/a425228.pdf.

Axe, David. "That Time an Air Force F-16 and an Army Missile Battery Fought Each Other." *Medium.* July 5, 2014. https://medium.com/war-is-boring/that-time-an-air-force-f-16-and-an-army-missile-battery-fought-each-other-bb89d7d03b7d.

Beer, Jenay, A.D. Fisk, and W.A. Rogers. “Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction,” *Journal of Human-Robot Interaction* 3, no. 2 (2014): 74-99. https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1127&context=csce_facpub

Bensahel, Nora. "Darker Shades of Gray: Why Gray Zone Conflicts Will Become More Frequent and Complex." *Foreign Policy Research Institute.* February 13, 2017. https://www.fpri.org/article/2017/02/darker-shades-gray-gray-zone-conflicts-will-become-frequent-complex/.

Brick, Joe. "The Military Profession: Law, Ethics, and the Profession of Arms." In *Redefining the Modern Military: The Intersection of Profession and Ethics*, by Nathan K. Finney and Tyrell O. Mayfield, 22-35. Annapolis: Naval Institute Press, 2018.

Brunnstrom, David. "China Installs Cruise Missiles on South China Sea Outpost." *Reuters.* May 02, 2018. https://www.reuters.com/article/us-southchinasea-china-missiles/china-installs-cruise-missiles-on-south-china-sea-outposts-cnbc-idUSKBN1I336G.

Bussby, Mattha and Anthony Cuthbertson. “Killer Robots Ban Blocked by US and Russia at UN Meeting.” *The Independent*. September 3, 2018. https://www.independent.co.uk/life-style/gadgets-and-tech/news/killer-robots-un-meeting-autonomous-weapons-systems-campaigners-dismayed-a8519511.html.

Clark, Ben. “Proportionality in Armed Conflicts: A Principle in Need of Clarification?” *Journal of International Humanitarian Legal Studies* 3, no. 1 (2012): 73-123.

Clausewitz, Carl Von. *On War.* Edited by Michael Howard and Peter Paret. Princeton: Princeton University Press, 1976.

Crist, David. *The Twilight War: The Secret History of America's 30-Year Conflict with Iran.* New York: Penguin Books, 2012.

Crowe, ADM William J. "Formal Investigation into the Circumstances Surrounding the Attack on the USS STARK (FFG-31) on 17 May 1987." Unpublished Report, Office of the Chairman of the Joint Chiefs of Staff, September 3, 1987. https://www.jag.navy.mil/library/investigations/USS%20STARK%20BASIC.pdf.

Defense Advanced Research Projects Agency (DARPA). "Accelerating the Exploration of Promising Artificial Intelligence Concepts." July 20, 2018. https://www.darpa.mil/news-events/2018-07-20a.

DARPA. "DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies." September 7, 2018. https://www.darpa.mil/news-events/2018-09-07.

_____. "Training AI to Win a Dogfight." May 8, 2019. https://www.darpa.mil/news-events/2019-05-08.

Defense Science Board. *Report of the Defense Science Board Task Force on Patriot System Performance - Report Summary.* Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2005. https://www.hsdl.org/?view&did=454598.

Denny, H.M. "Professionals Know When to Break the Rules." In *Redefining the Modern Military: The Intersection of Profession and Ethics*, by Nathan K. Finney and Tyrell O. Mayfield, 53-69. Annapolis: Naval Institute Press, 2018.

Deparment of Defense. *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance our Security and Prosperity.* Washington, DC: DoD, February 12, 2018. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

Department of Defense. *Autonomy in Weapon Systems.* Department of Defense Directive (DODD) 3000.09, Washington, DC: DoD, November 21, 2012. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

Erwin, Sandra. "Too Much Information, Not Enough Intelligence," *National Defense Magazine*. May 1 2012. http://www.nationaldefensemagazine.org/articles/2012/5/1/2012may-too-much-information-not-enough-intelligence.

Everett, Jim A.C., M.J. Crockett, and David A. Pizarro. "Inference of Trustworthiness from Intuitive Moral Judgments." *Jounal of Experimental Psychology General* 145, no 6 (June 2016): 772-787.

Flanagan, James D., and William N. Sweet. "AEGIS: Advanced Surface Missile System." *Johns Hopkins APL Technical Digest* 4, no. 4 (1981): 243-245. https://www.jhuapl.edu/techdigest/views/pdfs/V02_N4_1981/V2_N4_1981_Flanagan_Advanced.pdf.

Frederick, Bryan, and David E. Johnson. "The Continued Evolution of U.S. Law of Armed Conflict Implementation: Implications for the U.S. Military." Santa Monica: RAND, 2015. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1100/RR1122/RAND_RR1122.pdf.

Fogarty, RADM William M. "Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988." Unpublished Report, Office of the Chairman of the Joint Chiefs of Staff, August 19, 1988. https://www.jag.navy.mil/library/investigations/VINCENNES%20INV.pdf.

Freund, Eleanor. "Freedom of Navigation in the South China Sea: A Practical Guide." *Harvard Kennedy School.* June, 2017. https://www.belfercenter.org/publication/freedom-navigation-south-china-sea-practical-guide.

Gersh, John R. "Doctrinal Automation in Naval Combat Systems: The Experience and the Future." *Naval Engineers Journal* 99, no. 3 (May 1987): 74-79.

Global Security. "AEGIS Weapons System Mk 7." Accessed April 22, 2019. https://www.globalsecurity.org/military/systems/ship/systems/aegis-core.htm.

Green, Marc. "Night Vision." *Marc Green PhD Human Factors.* Accessed May 25, 2019. https://www.visualexpert.com/Resources/nightvision.html.

Green, Michael, Kathleen Hicks, Zach Cooper, John Schaus, and Jake Douglas. "Countering Coercion in Maritime Asia: The Theory and Practice of Gray Zone Deterrence." Washignton, DC: Center for Strategic and International Affairs (CSIS), May 2017. https://csis-prod.s3.amazonaws.com/s3fs-public/publication/170505_GreenM_CounteringCoercionAsia_Web.pdf?OnoJXfWb4A5gw_n6G.8azgEd8zRIM4wq.

Group of Governmental Experts on Lethal Autonomous Weapons Systems. “Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.” Working Papers of United Nations General Assembly, August 28, 2018. https://www.unog.ch/80256EDD006B8954/(httpAssets)/D1A2BA4B7B71D29FC12582F6004386EF/%24file/2018_GGE+LAWS_August_Working+Paper_US.pdf.

Gunning, David. "Explainable Artificial Intelligence." DARPA. Accessed May 19, 2019. https://www.darpa.mil/program/explainable-artificial-intelligence.

Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow.* New York: HarperCollins Publishers, 2017.

Horsburgh, H.J.N. "The Ethics of Trust." *The Philosophical Quarterly* 10, no. 40 (October 1960): 343-354.

International Committee of the Red Cross (ICRC), *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, 1125 UNTS 3, 8 June 1977. Accessed 30 May 2019. https://ihl-databases.icrc.org/ihl/WebART/470-750061?OpenDocument.

Johnson, Rebecca. "Ethical Requirements of the Profession: Obligations of the Profession, the Professional, and the Client." In *Redefining the Modern Military: The Intersection of Profession and Ethics*, by Nathan K. Finney and Tyrell O. Mayfield, 86-100. Annapolis: Naval Institute Press, 2018.

Joint Chiefs of Staff. *Joint Targeting*. Joint Publication (JP) 3-60. Washington, DC: Joint Chiefs of Staff, 28 September 2018.

Jonas, Hans. "Toward a Philosophy of Technology." *The Hastings Center Report* 9, no. 1 (1979): 11-25.

Kania, Elsa. "China's Artificial Intelligence Revolution." *The Diplomat*, July 27, 2017. https://thediplomat.com/2017/07/chinas-artificial-intelligence-revolution/.

LaGrone, Sam. "Destroyer that Protected U.S. Ships From Houthi Cruise Missiles Recognized as Best Atlantic Fleet Ship." *U.S. Naval Institute News,* October 18, 2017. https://news.usni.org/2017/10/18/destroyer-protected-u-s-ships-houthi-cruise-missiles-recognized-best-atlantic-fleet-ship.

Lange, Katie. "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are." *DoDLive.* March 30, 2016. http://www.dodlive.mil/2016/03/30/3rd-offset-strategy-101-what-it-is-what-the-tech-focuses-are/.

LCDR David Lee, JAGC, USN. ed. *Law of Armed Conflict Deskbook 2015*, 5th ed. Charlottesville, VA: U.S. Army Judge Advocate General's Legal Center and School, 2015. http://www.log.gov/rr/frd/Military_Law/pdf/LOAC-Deskbook-2015.pdf.

Lieber, Francis. "General Orders No. 100: The Lieber Code: Instruction for the Government of Armies of the United States in the Field." Yale University. Accessed April 22, 2019. http://avalon.law.yale.edu/19th_century/lieber.asp.

Lindemann, Michael. Response to "Doctrinal Automation in Naval Combat Systems." *Naval Engineers Journal* 99, no. 4 (July 1987): 108-109.

Linnan, David K. "Iran Air Flight 65 and Beyond: Free Passage, Mistaken Self-Defense, and State Responsibility." *Yale Journal of International Law* 16, no. 2 (1991): 245-389. http://digitalcommons.law.yale.edu/yjil/vol16/iss2/2.

Lucas, George. *Military Ethics: What Everyone Needs to Know*. New York: Oxford University Press, 2016.

Macdonald, Julia, and Jacquelyn Schneider. "Trust, Confidence, and the Future of Warfare." *War on the Rocks.* February 5, 2018. https://warontherocks.com/2018/02/trust-confidence-future-warfare/.

Magnetpraetorian. "US Missile Shoot Down - Iran Air Flight 655 Documentary." Video, 41:45. August 16, 2016. Accessed May 26, 2019. https://www.youtube.com/watch?v=1RJnumxuHwY.

Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. "Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents." In *Proceedings of the Tenth Annual AMC/IEEE International Conference on Human-Robot Interaction.* (March 2015): 117-124.

Malviya, Vishnu. "Five Best Books to Learn About Artificial Intelligence." *Technotification.* February 22, 2019. https://www.technotification.com/2019/02/best-books-for-artificial-intelligence.html.

McCarthy, J., M.L. Minsky, N. Rochester, and C.E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." Unpublished Research Proposal, Dartmouth College, August 31, 1955. https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

McDermott, Rose. *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy.* Ann Arbor: University of Michigan Press, 1998.

Missile Defense Advocacy Alliance. "AN/SPY-1 Radar." December, 2018. http://missiledefenseadvocacy.org/missile-defense-systems-2/missile-defense-systems/u-s-deployed-sensor-systems/anspy-1-radar/.

Navy. *U.S. Navy Fact File: AEGIS Weapons System.* Accessed April 09, 2019. https://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2.

Nothwang, William D., Ryan M. Robinson, Samuel A. Burden, Michael J. McCourt, and J. Willard Curtis. “The Human Should be Part of the Control Loop?” Unpublished Research, Office of the Secretary of Defense Autonomy Research Pilot Initiative, 2016. http://faculty.washington.edu/sburden/_papers/NothwangRobinson2016resil.pdf.

O’Connor, Jennifer. “Applying the Law of Targeting to the Modern Battlefield,” Department of Defense, Washington, DC. Last modified March 24, 2019.

https://dod.defense.gov/Portals/1/Documents/pubs/Applying-the-Law-of-Targeting-to-the-Modern-Battlefield.pdf.

O'Rourke, Ronald. *Navy Aegis Ballistic Missile Defense (BMD) Program: Background and Issues for Congress.* Washington, DC: CRS, April 2019. https://crsreports.congress.gov/product/pdf/RL/RL33745.

Pillar, Charles. "Vaunted Patriot Missile has a "Friendly Fire" Failing." *L.A. Times.* April 21, 2003. https://www.latimes.com/archives/la-xpm-2003-apr-21-war-patriot21-story.html.

Plato. "Theaetetus." *Plato in Twelve Volumes.* Translated by Harold N. Fowler. Cambridge, MA: Harvard University Press, 1921. http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0172%3Atext%3DTheat.%3Apage%3D152.

Polonski, Vyacheslov. "People Don't Trust AI--Here's How We Can Change That." *Scientific American*. January 10, 2018. https://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/?redirect=1.

Richardson, ADM John. "Countering Coercion in Maritime Asia." Remarks, Washington, DC, 2017. Center for Strategic and International Studies (CSIS). https://www.csis.org/analysis/remarks-cno-adm-richardson.

Riggs, Robert. "The Patriot Flawed? Failure to Correct Problem Led to Friendly Fire Deaths." *CBS News.* February 19, 2004. https://www.cbsnews.com/news/the-patriot-flawed-19-02-2004/.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach.* London: Pearson Education, 2010.

Sartor, Giovanni. "Doing Justice to Rights and Values: Teleological Reasoning and Proportionality." *Artificial Intelligence Law* 18, (2010): 175-215.

Saxon, Dan. 2014. "A Human Touch: Autonomous Weapons, Directive 3000.09, and the 'Appropriate Levels of Human Judgment over the Use of Force'." *Georgetown Journal of International Affairs* 15, no. 2 (Summer/Fall 2014): 100-109.

Scharre, Paul. *Army of None.* New York: W.W. Norton & Company, 2018.

Shaw, William H. *Utilitarianism and the Ethics of War.* New York: Routledge, 2016.

Singer, Peter, and August Cole. *Ghost Fleet.* New York: Houghton Mifflin Harcourt Publishing, 2015.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Vintage Books, 2017.

Tomes, Robert. "The Cold War Offset Strategy: Origins and Relevance." *War on the Rocks.* November 6, 2014. https://warontherocks.com/2014/11/the-cold-war-offset-strategy-origins-and-relevance/.

Tousley, Bradford. Quoted in Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*. New York: W.W. Norton & Company, 2018.

Tucker, Patrick. "US Military Testing Whether Human Pilots Can Trust Robot Wingmen in a Dogfight," *Defense One*. May 7, 2019. https://www.defenseone.com/technology/2019/05/us-military-testing-whether-human-pilots-can-trust-robot-wingmen-dogfight/156817/.

Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (September 27, 1974): 1124-1131.

Walzer, Michael. *Just and Unjust Wars.* New York: Perseus Books Group, 1977.

Work, Deputy Secretary Robert. "Remarks by Deputy Secretary Work on Third Offset Strategy: Delivered Brussels, Belgium." Brussels, April 2016. Department of Defense. Last modified April 28, 2016. https://dod.defense.gov/News/Speeches/Speech-View/Article/753482/remarks-by-d%20eputy-secretary-work-on-third-offset-strategy/.

Zhao, Huijie, Zheng Ji, Jianrong Gu, and Yansong Li. "Target Detection over the Diurnal Cycle Using a Multispectral Infrared Sensor." *Sensors (Basel)* 17, no. 56 (2017): 1-16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5298629/pdf/sensors-17-00056.pdf.