2022-02

Artificial Intelligence: Too Fragile to Fight?

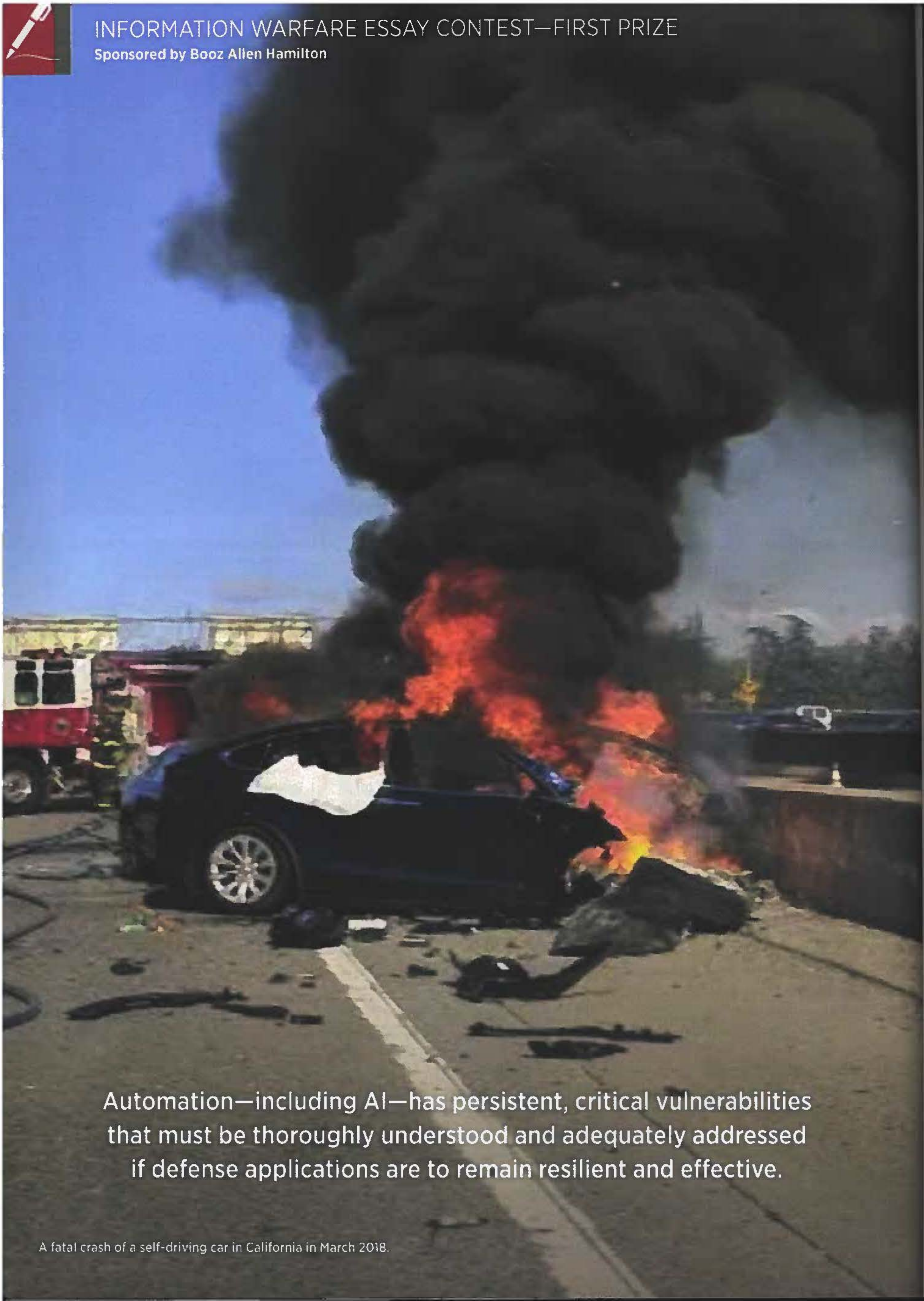
Jatho, Edgar; Kroll, Joshua A.

U.S. Naval Institute

Jatho, Edgar; Fox, Collin; Kroll, Joshua A. "Artificial Intelligence: Too Fragile to Fight?" Proceedings (US Naval Institute), February 2022.
http://hdl.handle.net/10945/68820

INFORMATION WARFARE ESSAY CONTEST—FIRST PRIZE

Sponsored by Booz Allen Hamilton



Automation—including AI—has persistent, critical vulnerabilities that must be thoroughly understood and adequately addressed if defense applications are to remain resilient and effective.

A fatal crash of a self-driving car in California in March 2018.

ASSOCIATED PRESS

ARTIFICIAL INTELLIGENCE

TOO FRAGILE TO FIGHT?

BY COMMANDER EDGAR JATHO, U.S. NAVY, AND JOSHUA A. KROLL

> *You can become utterly dependent on a new glamorous technology, be it cyber-space, artificial intelligence. . . It'll enable you. It'll move you forward. But does it create a potential achilles heel? Often it does.*[1]
>
> —Admiral James Stavridis

Artificial intelligence (AI) has become the technical focal point for advancing naval and Department of Defense (DoD) capabilities. Secretary of the Navy Carlos Del Toro listed AI first among his priorities for innovating U.S. naval forces. Chief of Naval Operations Admiral Michael Gilday listed it as his top priority during his Senate confirmation hearing.[2] This focus is appropriate: AI offers many promising breakthroughs in battlefield capability and agility in decision-making.

Yet, the proposed advances come with substantial risk: automation—including AI—has persistent, critical vulnerabilities that must be thoroughly understood and adequately addressed if defense applications are to remain resilient and effective. Current state-of-the-art AI systems undergirding these advances are surprisingly fragile—that is, easily deceived or broken or prone to mistakes in high-pressure use.

Machine learning (ML) and especially modern "deep learning" methods—the very methods driving the advances that make AI an important focus area today—are distinctly vulnerable to deception and perturbation.[3] Often, human-machine teaming is thought to be the solution to these issues, but such teaming itself is fraught and unexpectedly fragile in persistently problematic and counterintuitive ways.[4]

This foundation-level fragility is a potentially catastrophic flaw in warfighting systems. It undermines expectations, since new and seemingly capable systems appear to outperform existing technology under straightforward evaluation. However, failure modes in future applications are often invisible. Thus, AI proponents rightly claim major technological advances but often fail to adequately acknowledge the limitations of those advances. This in turn risks overreliance on technology that may fall significantly short of expectations.

Consider this quote from General Mike Murray, commanding general of Army Futures Command and a leader in DoD technology adoption, during a recent radio interview:

> I can't imagine an automated target recognition system not doing a better job than human memory can do. . . Say you had to have a 90 percent success rate on flash cards to qualify to sit in that gunner's seat. With the right training data and right training, I can't imagine not being able to get a system, an algorithm, to be able to do better than 90 percent in terms of saying this is the type of vehicle you are looking at and then allowing that human to make the decision on whether it is right or not and then pull the trigger.[5]

This statement reflects a failure of imagination about the limits of AI and the difficulties in the hand-off between humans and automation. The claims of "success rate" are derived from limited-scope experiments and do not provide a dependable case that such systems are ready for deployment, even on a trial basis. Instead, a careful examination of technological reality is needed. Such an examination must consider pitfalls and lessons learned from the past half century of implementing automation in large critical-domain systems (such as aviation, manufacturing, and industrial control systems). There are many challenges that arise in such systems, and a stronger case for adoption comes with understanding these inherent issues.

MISPLACED OPTIMISM

Current AI claims are often wildly optimistic.[6] Such claims inflate expectations of what this technology can do, risking disillusionment when the technology fails to deliver. AI is not a cure-all that applies in all cases, nor a product one can simply buy and implement. Rather, AI is a set of techniques that reshape problems and their solutions. Dependable application of AI to military or national security problems must rest on concrete foundations to justify confidence in the system.[7] Limitations must be identified to be overcome, and the military must not rush forward into new technology on incomplete arguments that ignore fundamental technical reality. Otherwise, it may find itself depending on brittle tools not up to the task of actual warfare.

A CURE WORSE THAN THE DISEASE?

In military operations, new technologies must be carefully evaluated against the standard of whether adopting them creates unknown and possibly more insidious problems than it solves. For large, complex, and "wicked" problems, it is not always or even often the case that "any solution is better than no solution." Rather, interventions frequently create new problems, and proponents of novel approaches have an attendant responsibility to justify confidence in them.

To that end, several known deficiencies in current AI systems are discussed below to determine what a case for trustworthy interventions would require. We use General Murray's notional AI-based targeting system as a running example because it is a setting in which much attention is given in research, development, and policy circles.[8]

INCOMMENSURABLE GOALS

Human recognition and a target-recognition algorithm are neither equivalent nor directly comparable. They perform different tasks in different ways and must be measured for success against different metrics.

Human recognition in a targeting task describes not just identifying a target but also discerning why a portion of a scene might be targetable. Humans understand concepts and can generalize their observations beyond the situation, loosely gauge uncertainty in their assessments about target identification, and interpret novel scenarios with only minimal confusion. For this reason, human vision and discernment do far more than a simple target-recognition flash-card test can measure.

An AI system's target recognition is vacant by comparison. An automated vision-based classification system does far less than the term "recognition" implies, a term that implicitly anthropomorphizes algorithmic systems that simply interpret and repeat known patterns. Such systems cannot understand the reasons that targets should be selected nor generalize beyond the specific patterns they are programmed to handle. Rather, these systems apply patterns, which are either programmed or extracted from data analysis. In a scenario that has never been encountered before, it is possible that no known pattern applies. AI systems will give guidance nonetheless—knowledgeless, baseless guidance.

In the real world of varying environments, degraded equipment, or in which deliberate evasion and deception are expected, performance on image recognition alone does not describe performance for the extended task of target recognition.[9] Humans are far superior at dealing with image distortions (e.g., dirt or rain on the camera lens, electrical noise in a video feed, dropped portions of an image from unreliable communications). Models trained on specific image distortions can approach or exceed human performance on that particular distortion,

U.S. NAVY (BRANDON ROBERSON)

A Boeing unmanned MQ-25 aircraft is given operating directions on the USS *George H. W. Bush* (CVN-77) flight deck in December 2021. In military operations, new unmanned technologies must be carefully evaluated against the standard of whether adopting them creates unknown and possibly more insidious problems than they solve.

but the improvement does not translate to better performance on any other type of distortion.

Although it may be true that image recognition models can "outperform" humans on simple flash-card style tests, equating human and algorithmic performance in target selection and discernment using laboratory data or in an operational test scenario, as General Murray referred to, implies that performance on these tasks is comparable. This is simply false. The work being done in each case is not the same, and the reliability of generated answers is vastly different. Raw performance is misleading, and relying on it could lead to dangerous situations.

ADVERSARIAL DECEPTION AND AUTOMATION BIAS

The current best-performing AI approaches, based on deep neural network machine learning, can seem to outperform humans on simple flash-card style qualification tests. This performance comes at a high cost, however: Such models over-learn the details of the evaluation criteria instead of general rules that apply to cases beyond the test. A particularly noteworthy example is the problem of "adversarial examples," situations designed by an adversary to confuse the technology as much as possible.[10] Some researchers have suggested that AI's susceptibility to adversarial deception may be an unavoidable characteristic of the methods used.[11] This is not a new problem for warfighting—camouflage exists in nature and has been practiced in an organized fashion in military units for hundreds if not thousands of years. Rather, to use AI effectively, the military must be aware of the extent to which deception can cause misbehavior and build the attendant doctrine and surrounding systems such that the AI-supported decisions remain robust even when adversaries attempt to influence them.

One might imagine that the problem of machine fragility can be resolved by keeping a human in the decision loop—that is, an AI system recommends actions to a human or is tightly supervised by a human, such that the human is in control of the outcome. Unfortunately, human-machine teams often prove to be fragile as well. When guided by automation, humans can become confused about the state of the automation and the appropriate control actions to take. In July 1988, the USS *Vincennes* (CG-49) accidentally shot down an Iranian civilian airliner departing its stopover at the Bandar Abbas International Airport after the ship's Aegis system reused a tracking identifier previously assigned to the civilian airliner for a fighter jet far from the ship's position. When asked to describe the activity of the contact using the old tracking identifier, the human operator correctly indicated that it was a fighter that was descending, information that led (together with the track of the civilian airliner toward the ship) to a decision to fire on the track identified by the old identifier.[12] Although automation has improved, today human-machine team fragility has accounted for recent crashes of highly automated cars such as Teslas, the at-sea collision of the USS *John S. McCain* (DDG-56) in 2017, and in the loss of Air France Flight 447 over the Atlantic in 2009.

This underscores the problem of mode confusion between humans and machines, which can be exacerbated when information moves in complex systems or

U.S. NAVY (ARTHUR ROSEN)

An operations specialist monitors surface contacts from the combat information center on board the USS *John S. McCain* (DDG-56) in the East China Sea. Despite their fragility, human-machine teams can massively outperform either humans or machines, so long as the right functions are assigned to each.

is presented with poor human factors. A related problem is automation dependency, in which humans fail to seek out information that would contradict machine solutions. In both cases, assessing how well human-machine teams perform in context is critical—whether the goal is to improve performance on average or in specific, difficult situations.

It might be argued that high overall performance or certification for operation in particular applications negates these concerns. But this also is an oversimplified view. Consider the targeting scenario from General Murray again, refining the hypothetical performance numbers: Suppose the system has a 98 percent accuracy, but a trained human has only an 88 percent accuracy on the same set of test scenarios. For human operators on the battlefield, when bullets and missiles are flying and lives hang in the balance, will the operators question the system's claims, or will they simply pull the trigger? Can operators trust that because the machine is better in aggregate it translates to better performance now, in their specific situation?

AUTOMATION PARADOX

As tasks are automated away from daily practice, human operators suffer what is referred to as de-skilling.[13] So operators of General Murray's hypothetical tank system, while required to "catch" the mistakes of the system, are unfortunately less qualified to do so because they no longer are routinely required to perform the task unaided. For instance, consider what smartphone GPS-based navigation has done for the average person's wayfinding skills: A once-routine task is now unmanageable for many. This phenomenon affects professionals such as pilots and even bridge watch teams.

Despite their fragility, human-machine teams can also massively *outperform* either humans or machines, so long as the right functions are assigned to each part and the appropriate affordances are made by humans for machines and machines for humans. Consider the game of "cyborg chess" (or "advanced chess"), in which human players use computer decision aids in selecting their moves. In tournaments, even chess players who are weak when unaided can play at a level that exceeds the world's top grandmasters and the world's top computer chess programs. Thus, human-machine integration and a focus on the processes surrounding automation can be far more impactful than human skill or intelligence alone.[14]

The military must not approach AI applications as self-contained artificial minds providing clear outputs fusing all features. Instead, AI must be an extension of human intelligence and organizational capability. AI is not an independent agent, but a more capable tool, applied to specific aspects of existing operations.

MULTISENSOR HOPES

If a vision-based system is fragile, perhaps a system that fuses many types of sensors is better? The logical extension to vision-based systems is multiple sensor-data inputs to enhance an AI system's capability to find, fix, track, and target reliably. This approach is currently under evaluation in the Scarlet Dragon exercise.[15] Cross-cueing inputs from different domains (e.g., visual and electronic signatures) is analogous to human multisensory perception. For example, when what a human hears does not match associated visual stimuli, it raises suspicions and draws scrutiny that may uncover the deception.

It is an open question, however, whether this approach improves robustness against adversarial manipulation of AI systems. Each sensor's data input to an automated tool is still subject to the same adversarial techniques. The added complexity induces a trade-off. On the one hand, multiple sensors complicate the adversary's challenge in deceiving the system. On the other, increasing the number of input elements and the complexity of the features in a model also leads in a mathematically inexorable way to a greater potential for adversarial manipulation (because the number of possible deception approaches increases faster than the number of valid inputs). More study is needed to find the optimal trade-off. However, a move to sensing in multiple domains certainly does not foreclose the possibility of deception or even any specific avenue.[16]

The above discussion notwithstanding, there is no denying the urgent need to move forward with AI in Navy and wider DoD applications. However, the warfighters' eyes must be wide open. They must be extremely judicious about when, where, and how they employ these technologies. In support of such care, they should consider the following three principles for judicious and responsible deployment of AI systems in DoD applications:

- Absent strong evidence, remain skeptical of claims that these systems work as well as reported. Training datasets, environment, test conditions, and assumptions all have outsized effects on results. Practical translation of industry findings to warfighting requirements is not straightforward.
- AI systems must be deployed only with adequate technical and sociotechnical safety nets in place. Overcoming environmental and adversarial perturbations are difficult, unsolved problems. Because AI operates based on patterns (programmed or extracted from data), its ability to operate when those patterns do not hold is inherently limited.
- Human-machine teams must be tested and measured as a system together. Humans and machines are good at different parts of any problem. Allocating functions and composing these capabilities is not straightforward, but often counterintuitive. Careful assessment of the entire system is required to ground any claims about trustworthiness or suitability for an application.

AI succeeds best when it solves a clear, carefully defined problem of limited scope, supporting the existing work of warfighters or the DoD enterprise. In a world in which U.S. leaders warn of a risk of losing competitive military-technical advantage if the nation does not adopt the newest technologies, it is imperative that naval leaders understand the inherent limitations of AI so its adoption in critical warfighting capabilities will not incur catastrophic vulnerabilities at their heart.

[1] ADM James Stavridis, USN (Ret.), and John Arquilla, "Weapons of Mass Disruption: A Conversation on the Future Force, Geopolitics and Leadership," The Naval Postgraduate School's Secretary of the Navy Guest Lecture, 12 October 2021, www.youtube.com/watch?v=p1XQfNv2PXU.

[2] U.S. Department of Defense, ADM Michael Gilday, USN, Confirmation testimony to the Senate Armed Services Committee, 31 July 2019, www.defense.gov/Multimedia/Photos/igphoto/2002165109/.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," arXiv preprint arXiv:1312.6199, 2013, arxiv.org/abs/1312.6199; Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry, "On Adaptive Attacks to Adversarial Example Defenses," presented at the 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020, proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf.

[4] Mary L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," in *Decision Making in Aviation*, 289–94.

[5] GEN Mike Murray, USA, in interview: Meghna Chakrabarti, John M. Murray, Patrick Tucker, Heather Roff, Gilman Louie, and Mikel Rodriguez, "Understanding the AI Warfare and Its Implications," WBUR On Point, July 2021, www.wbur.org/onpoint/2021/07/29/understanding-the-ai-warfare-and-ethics.

[6] One example of this implicit hyperbole is represented by the chart found on page 343 in the 2019 Economic Report of the President titled "Error Rate of Image Classification by Artificial Intelligence and Humans, 2010–17." It shows "human classification" error rate charted next to a machine classification error rate. The graph and text present as fact that machine vision surpassed human image classification capabilities in 2015. Careful consideration of this claim and examination of the referenced research and even current state of the art research reveals this particular development still remains a distant as-yet unreached milestone.

[7] One successful approach to justifying confidence in a desired system property in complex systems can be borrowed from methods adopted by designers of aircraft and nuclear power plants to ensure safety, i.e., assurance cases.

[8] Patrick Tucker, "How Well Can AI Pick Targets from Satellite Photos? Army Test Aims to Find Out," *Defense One*, 6 October 2021.

[9] Even with models trained on standard reference datasets, there is a well documented reduction in performance when said models are tested on disjointed sets of the same provenance. In contrast, human performance does not suffer this defect in performance testing between data sets. Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt, "Evaluating Machine Accuracy on ImageNet," delivered at the Proceedings of the 37th International Conference on Machine Learning, *Proceedings of Machine Learning Research* 119 (2020): 8634–44, proceedings.mlr.press/v119/shankar20c.html; Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann, "Generalization in Humans and Deep Neural Networks," presented at the 32nd Conference on Neural Information Processing Systems, Montreal, Canada, 2018, proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," arXiv preprint arXiv:1312.6199, 2013, arxiv.org/abs/1312.6199

[11] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein, "Are Adversarial Examples Inevitable?" International Conference on Learning Representations, 2019, arXiv preprint arXiv: 1809.02104, arxiv.org/pdf/1809.02104.pdf

[12] Craig W. Fisher and Bruce R. Kingma, "Criticality of Data Quality as Exemplified in Two Disasters," *Information & Management* 39, 2 (December 2001): 109–16.

[13] Lisanne Bainbridge, "Ironies of Automation," *Automatica* 19, 6 (November 1983), 775–79.

[14] Pontus Wärnestål, "Why Human-Centered Design Is Critical to AI-Driven Services, *Inuse*, 9 September 2019, www.inuse.se/read/why-human-centered-design-critical-ai-driven-services-e242a8067af.

[15] Patrick Tucker, "How Well Can AI Pick Targets from Satellite Photos? Army Test Aims to Find Out."

[16] As a further complication, when the AI has access to more sources of data and sensor types than the human can quickly aggregate, dependably consume, and interpret (many times the very reason we need to adopt AI), this further compounds automation bias, ensuring the human operator only feels justified in arriving at the decision suggested by the system. As it has much more information at its instantaneous disposal, it will naturally come to be seen as superior given the breadth of information with which it arrives at a decision.

■ **COMMANDER JATHO** is a member of the Navy's permanent military professor community. He is pursuing a PhD in computer science at the Naval Postgraduate School in Monterey, where his research focuses on trustworthy AI and defending deep neural networks from adversarial attacks and deception. Prior to May 2020, he served in the cryptologic warfare community. Tours included Navy Cyber Defense Operations Command as the defensive cyber operations afloat department head, Carrier Strike Group 10 as cryptologic resource coordinator, and National Security Agency as deputy chief, special access program central office/special technical operations.

■ **DR. KROLL** is an assistant professor of computer science at the Naval Postgraduate School in Monterey, California. He received a PhD in computer science from Princeton University with a focus on cybersecurity and technology policy. Previously, he was a researcher at UC Berkeley and worked at the internet security and performance company Cloudflare. His research focuses on enabling the responsible and trustworthy use of new computational technologies.