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.
https://hdl.handle.net/10945/68820

Artificial Intelligence: Too fragile to fight?

CDR EDGAR W. JATHO, USN, Naval Postgraduate School, USA
JOSHUA A. KROLL, PHD, Naval Postgraduate School, USA

"*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.*"
Admiral James Stavridis, USN (Ret.) [8]

1 INTRODUCTION

Artificial Intelligence (AI) has become the technical focal point for advancing Naval and Department of Defense capabilities. Secretary of the Navy Del Toro listed AI first among his priorities for innovating US Naval Forces. CNO Admiral Gilday listed it as his top priority during Senate confirmation [10]. 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, broken or prone to mistakes in high-pressure use.

Machine learning (ML) and other 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 [9, 11]. 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 counter-intuitive ways [3].

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

Consider this quote from a leader in DoD technology adoption during a recent major-media 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% 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% 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.
>
> – General John Murray, CG, Army Futures Command Interview with *On Point* July, 2021 [2].

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 test basis. Instead, we need a careful examination of technological reality. Such an examination would consider pitfalls and lessons learned from the past half century of

implementing automation in large critical-domain systems.[1] There are many challenges that arise in such systems and a stronger case for adoption comes with an understanding these inherent issues.

Current AI claims are often wildly optimistic.[2] Such claims inflate our 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, an argument justifying confidence in the system.[3] Limitations must be identified to be overcome, and we must not rest our rush forward into new technology on incomplete arguments that ignore fundamental technical reality. Otherwise, we may find ourselves dependent on brittle tools not up to the task of actual warfare.

2 A CURE WORSE THAN ITS DISEASE?

In critical applications like military operations, we must carefully evaluate new technologies against the standard of whether adopting them creates unknown, possibly more insidious problems than they solve. 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, we summarize a number of known deficiencies in current AI systems, in support of outlining what a case for trustworthy interventions would require. In our analysis, we use General Murray's notional AI-based targeting system as a running example because it is a setting where there is much attention in research, development, and policy circles [12].

2.1 Incommensurable Goals

To begin our analysis, *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 and recognizing a target, but discerning and reasoning about why a portion of a scene might be targetable. Humans understand concepts, can generalize their observations beyond the particular situation, loosely gauge uncertainty in their assessments about target identification and can interpret novel scenarios not previously encountered with only minimal confusion. For this reason, human eyes and discernment do far more than is measured by a simple target-recognition flash card test.

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

[1] Such as aviation, manufacturing, and industrial control systems.

[2] One example of this implicit hyperbole is represented by the chart found on page 343 in the 2019 U.S. Economic Report of The President titled "Error Rate of Image Classification by Artificial Intelligence and Humans, 2010–17". It represents a constant 5% for humans charted next to a line that passes well below the human line beyond 2015.

[3] 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.

In the real world of varying environments, degraded equipment, or where deliberate evasion and deception are expected, performance on image recognition alone does not describe performance for the extended task of target recognition.[4] 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, but the improvement does not translate to better performance on any other type of distortion [5].

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 in Gen. Murray's quote, 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.

2.2 Adversarial Deception: A Persistent Problem

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 [9]. Some researchers have suggested that the susceptibility of AI to adversarial deception may be an unavoidable characteristic of the methods used [6]. 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, we must be aware of the extent to which deception can cause misbehavior and must build the attendant doctrine and surrounding systems such that the decisions supported by the AI remain robust even when adversaries attempt to influence them.

2.3 Automation Bias

One might imagine that the problem of machine fragility can be resolved by keeping a human "in or on" 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 actually 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 as a result. Examples abound: in July 1988, the *USS Vicennes* accidentally shot down a civilian airliner departing its stopover at the Bandar Abbas International Airport after the ship's Aegis system re-used a tracking identifier assigned to the plane 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, leading to the disaster [4]. 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 is presented with poor human factors. A related problem is *automation*

[4]Even with models trained on standard reference datasets, there is a well documented reduction in performance when said models are tested on disjoint sets of the same provenance [7]. In contrast, human performance does not suffer this defect in performance testing between data sets [5].

dependency, where 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, understanding 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. This is also an oversimplified view. Let us consider the targeting scenario from Gen. Murray again, refining the hypothetical performance numbers: suppose the system has a 98% accuracy, but a trained human only has an accuracy of 88% on the same set of test scenarios. For a human operator on the battlefield, when bullets and missiles are flying and the lives of their countrymen hang in the balance, will the operator question the system's claims or will they simply pull the trigger? Can the operator trust that the machine's better *in aggregate* translates to better performance now, in their specific situation?

2.4 Automation Paradox

As tasks are automated away from daily practice, human operators suffer what is referred to as de-skilling [1]. So operators of Gen. Murray's hypothetical tank system, while required to "catch" the mistakes of the system, are unfortunately less qualified to do so because they are no longer routinely required to perform the task unaided. Everyday examples abound, for instance, consider what smartphone GPS-based navigation has done for the average person's wayfinding skills: a once-routine task is now untenable 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 affordances made by humans for machines and machines for humans are appropriate. 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 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 the human's skill or intelligence [13].

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

2.5 Multi-Sensor Hopes

If a vision-based system is fragile, perhaps a system which fuses many types of sensors is better? The logical extension to vision based systems is the use of 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 [12]. Cross cuing inputs from different domains (e.g., visual and electronic signatures) is analogous to our own multi-sensory perception. For example, when what a human hears does not match associated visual stimuli, it raises suspicions and results in scrutiny that may uncover the deception.

It is an open question, however, whether this approach actually 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 tradeoff: on the one hand, multiple sensors complicate the adversary's challenge in deceiving the system; on the other hand, 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). To find the optimal trade-off, more study of the problem space is needed. However, a move to sensing in multiple domains certainly does not foreclose the possibility of deception or even any specific avenue.[5]

3 CONCLUSION

The above discussion notwithstanding, there is no denying the urgent need to move forward with AI in Navy and DoD applications. However, the warfighter's eyes must be wide-open. We must be extremely judicious about when, where and how we employ these technologies. In support of such care, we offer three principles for considering the judicious and responsible deployment of AI systems in DoD applications:

(1) 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.
(2) AI systems must only be deployed with adequate technical and socio-technical 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.
(3) Human-machine teams must be tested and measured as a system together. Humans and machines are good at different parts of any problem set. Allocating functions and composing these capabilities isn't straightforward, but often counter-intuitive. 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 where leaders warn of a risk of *losing* our competitive military-technical advantage if we don't adopt the newest technologies, it is imperative that Naval leaders understand the inherent limitations of AI so that its adoption in our critical warfighting capabilities will not incur catastrophic vulnerabilities at their heart.

REFERENCES

[1] Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (1983), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8
[2] Meghna Chakrabarti, John M Murray, Partick Tucker, Heather Roff, Gillman Louie, and Mikel Rodriguez. 2021. Understanding The AI Warfare And Its Implications. *On Point* (Jul 2021).
[3] Mary L Cummings. 2017. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation*. Routledge, 289–294.
[4] Craig W Fisher and Bruce R Kingma. 2001. Criticality of data quality as exemplified in two disasters. *Information & Management* 39, 2 (2001), 109–116.
[5] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. 2020. Generalisation in humans and deep neural networks. arXiv:1808.08750 [cs.CV]
[6] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. 2019. Are adversarial examples inevitable?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lWUoA9FQ
[7] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating Machine Accuracy on ImageNet. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 8634–8644. https://proceedings.mlr.press/v119/shankar20c.html

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

[8] James Stavridis and John Arquilla. 2021. Weapons of Mass Disruption: A Conversation on the Future Force, Geopolitics and Leadership. https://www.youtube.com/watch?v=p1XQfNv2PXU The Naval Postgraduate School's (NPS) Secretary of the Navy Guest Lecture (SGL) features retired U.S. Navy Adm. James G. Stavridis and NPS Distinguished Professor Emeritus John Arquilla discussing the topic in title.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[10] 78th Secretary of the Navy The Honorable Carlos Del Toro. 2021. One Navy-Marine Corps Team: Strategic Guidance From The Secretary of the Navy. https://media.defense.gov/2021/Oct/07/2002870427/-1/-1/0/SECNAV%20STRATEGIC%20GUIDANCE_100721.PDF

[11] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347* (2020).

[12] Patrick Tucker. 2021. How Well Can AI Pick Targets From Satellite Photos? Army Test Aims to Find Out. *Defense One* (October 2021). https://www.defenseone.com/technology/2021/10/how-well-can-ai-pick-targets-satellite-photos-army-test-aims-find-out/185916/

[13] Pontus Wärnestål. [n. d.]. Why Human-Centered Design is Critical to AI-Driven Services. ([n. d.]). https://medium.com/swlh/why-human-centered-design-is-critical-to-ai-driven-services-e242a8067af

NOTES