MAD IN AN AI FUTURE?

June 3, 2019

Joseph Johnson

Center for Global Security Research
LAWRENCE LIVERMORE NATIONAL LABORATORY

MAD in an AI Future?

Joseph Johnson[1]
Center for Global Security Research

Notwithstanding concerns to the contrary,[2] strategic nuclear deterrence, as understood in the context of mutually-assured destruction (MAD), is highly unlikely to be upset by advances in artificial intelligence (AI) in the foreseeable future. AI will improve the processes and systems that enable MAD and modern C^{4}ISR (command, control, communications, computers, intelligence, surveillance, and reconnaissance), but advances in the field are unlikely to reach the sophistication, accuracy, and resilience required to disrupt nuclear deterrence as understood and practiced since the end of World War II.

During the Cold War, MAD deterred nuclear aggression between the United States and Soviet Union by assuring that if either state attacked with nuclear weapons, both were certain that the attacked party could and would retaliate with nuclear weapons, resulting in the destruction of both states. “Assured” is the essential element of MAD—the targeted state must be known to possess survivable retaliatory nuclear forces. The Cold War saw the development and maturation of the modern nuclear triad, in which both sides deploy air, ground, and sea nuclear forces making a disarming first-strike practically impossible, all but assuring an unacceptable nuclear response.

If a state did develop and operationally deploy a credible “counterforce” capability to prevent retaliatory nuclear attacks, the premise of MAD and nuclear deterrence would be fundamentally disrupted. While the notion of an unassailable counterforce has been considered technologically infeasible, recent advances in imaging platforms and AI may put it back on the table.[3] Lieber and Press posit an increasing vulnerability of nuclear forces, owing to the improved accuracy of opposing nuclear delivery systems and a “revolution in remote sensing.” They warn that states facing “technologically advanced adversaries” will be particularly vulnerable, because “guidance systems, sensors, data processing, communication, artificial

[1] The author thanks Wes Spain and Zack Davis at the Center for Global Security Research, Lawrence Livermore National Laboratory for their contributions to this assessment.

[2] Groll, Elias, “How AI Could Destabilize Nuclear Deterrence,” *Foreign Policy*, April 24, 2018. Accessed August 8, 2018, https://foreignpolicy.com/2018/04/24/how-ai-could-destabilize-nuclear-deterrence/; Geist, Edward and Andrew J. Lohn, “How Might Artificial Intelligence Affect the Risk of Nuclear War?”. Santa Monica, CA: RAND Corporation, 2018. https://www.rand.org/pubs/perspectives/PE296.html.

[3] Lieber, Keir A. and Daryl G. Press, “The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence,” *International Security*, Vol. 41, No. 4 (2017).

intelligence, and a host of other products of the computer revolution continue to improve" and may be integrated to form a robust counterforce capability.[4]

While significant advances in guidance systems, sensors, data processing, and communications have shaped nuclear-weapon systems and counterforce capabilities, the role AI will ultimately play is less clear. AI may bring modest improvements in certain applications important to a credible counterforce capability—for example, in the finding, identifying, and precise locating of nuclear-weapon delivery platforms, especially mobile platforms such as missiles, aircraft, and submarines—but they are unlikely to fundamentally transform this part of the counterforce challenge. And because finding, identifying, and precisely locating adversarial nuclear-weapon delivery platforms will remain an unsolved problem, any improvement to a nation's confidence in its counterforce capability will be modest. MAD-based nuclear deterrence is likely to persist for the foreseeable future.

Limits of Artificial Intelligence for the Counterforce Challenge

The counterforce challenge is daunting. A nuclear state would need an extremely high degree of confidence that it could identify and preemptively destroy or disable all adversarial nuclear-weapon delivery systems capable of launching devastating retaliatory attacks. While a state may accept some risk associated with less than 100% certainty, if the intent of a counterforce attack is to disarm the adversary before it can attack with any nuclear weapons, then targeting certainty very near 100% is required.

Near-perfect performance on the part of machines is limited to highly predictable and controlled environments. The surveillance and precision-munitions technologies would hold up well in their respective roles of providing imagery and neutralizing targeted weapons, because the physics that governs these operations is well-defined. However, AI's role in the process, namely, the automated detection of weaponry, is based on an operation that is poorly understood— that of human vision. Quantum-computing pioneer David Deutsch proved that if the mechanics of sight were fully understood, they could be "emulated… on a general-purpose computer, provided it is given enough time and memory."[5] Without benefit of this insight, AI research has pursued diverse paths towards human-level intelligence.

The best results for object-detection come from the AI subfield of machine learning (ML), particularly deep-learning (DL) methods. Based on results sufficiently impressive to fan an AI reawakening, DL could improve identification, reconnaissance, and surveillance performance to

[4] Ibid., pp. 9-10.

5 Deutsch, David. "Creative Blocks", 2012, https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence.

increase the transparency of an adversary’s nuclear arsenal, particularly if interleaved with human analysis. However, a careful delineation of the constraints inherent to the idea of a counterforce shows that AI cannot reliably perform image-recognition tasks at the near-perfect level required; it will encounter a number of fundamental and immutable ML theoretical limits, rendering automated detection a weak link in any counterforce.

This analysis focuses on land-based mobile missiles only, deployed on transporter-erector-launchers (TELs) or railroads—a challenging counterforce target that is representative of many countries’ nuclear order of battle.[6]

AI is only as good as the data and information it was trained on and the data it is fed while operationally deployed. While advances in reconnaissance and surveillance systems have been impressive, the quantity and quality of data collected and processed still have limits.

- Effective automated object recognition requires two elements: images of the objects in question and information pinpointing the objects within the image and their classifications. The image and labels are together termed the “ground truth.” The difficulty of collecting images with intelligence assets weakens any guarantee of continuous and sufficient access to ground-truth images. Any lapse in this pipeline may result in undetected changes to the design of the arsenal.
- Lack of total access to the adversary’s nuclear arsenal limits the number of ground-truth images we can obtain. Even a free flow of information does not preclude the enemy from poisoning the data before we obtain it.[7]
- Due to the limitations of other platforms, satellites will remain heavily relied on for the detection, identification, classification, and location of mobile launch systems. Satellites alone can provide persistent, or near-persistent, imaging of adversary land-based systems. This dependency limits the quality of images, in terms of resolution and variety of angles, as well as the quantity collected for ground truth.
- The reconnaissance and surveillance systems required for credible counterforce will need to detect threats amid hundreds or thousands of vehicles simultaneously and recognize vehicles that are off road. Hence, AI will have to account for an extraordinary variety of terrains.
- Detection and identification is not a static problem. Adversaries will continue to develop the camouflage, movement, and concealment of mobile weapons and employ decoys.

[6] TELs are tracked or wheeled vehicles that move on- or off-road and are set up quickly to launch a ballistic missile.

[7] If our intelligence assets were strong enough with respect to a particular adversary that a continuous flow of ground-truth images were available, AI would not likely be necessary to track the adversary’s nuclear arsenal.

Data Challenges for Counterforce Mission

The detection and identification processes of AI-enabled object recognition start with imagery-collection systems inputting image data into our AI. According to Lieber and Press, current synthetic-aperture radar (SAR) provides clean, overhead images in 150-by-150 kilometer swaths, regardless of cloud cover or time of day.[8] Upon detecting a vehicle, a magnified image is input to the AI[9] system to determine the probability that it is a mobile launcher. The data quality for this task is poor for three reasons: it is imbalanced, a poor representation of a mobile weapon, and subject to the "curse of dimensionality."

Data Imbalance

While the quantity of data input to AI will be large, the AI system will have relatively few ground-truth images of TELs or rail-based launchers to train on. Since assembling a dataset of adversary mobile launchers is a manual process, the number of correct original images to learn from will be relatively few and may not represent all models. This dearth may have subtle effects on AI results. Because it must distinguish between mobile launchers and other moving objects, the AI must be trained on images of commercial and other military (non-TEL) vehicles, for which there is an abundance of data. If the number of non-TEL images is much greater than the number of TEL images, this data imbalance will incentivize the AI to increase its accuracy by rarely or never identifying a mobile launcher.[10]

Recent developments address the problem of imbalance by taking smaller datasets and expanding them by creating variations of each data point or "synthetic data".[11] We can take an image of a mobile launcher and warp it several different ways, thereby generating additional images for each image we possess. Training the AI on this additional data expands its concept of a mobile launcher, leading to the identification of TELs that would otherwise go unflagged. However, this method of increasing the number of correctly identified objects, or "true positives," comes at a cost. A portion of the synthesized images will, according to the AI, look similar to non-TEL vehicles.[12] The AI will classify these vehicles as mobile launchers,

[8] Lieber, Keir A. and Daryl G. Press, pp. 38-39.

[9] Most likely the AI would employ a neural network, in which case the neurons in the neural network would be picking out different features of the image.

[10] AI is trained on data we already have. The training process is a methodical tweaking of AI parameters to improve accuracy on the training data. Consider the case where 1% of our training images are of TELs. The AI can achieve 99% accuracy simply by classifying each image as a non-TEL. It is likely that the resulting parameters would be significantly different than those of AI that properly classifies the 1% TEL images. (The latter AI may or may not have high accuracy on non-TEL images.) Therefore, a tension will generally exist between training for high overall accuracy or high accuracy with respect to TEL images.

[11] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." *Advances in Neural Information Processing Systems*, (2014): 2672-2680 .

[12] All CV algorithms work the same in that they compute a "distance" between images where images that are close together are grouped into the same class. However, the groupings are rarely clean. Many examples appear to belong equally likely to multiple classes. In the case of recognizing mobile missiles, the images will fall along some sort of continuum of probabilities of being a TEL. See Domingos, Pedro. "A few useful things to know about machine learning." *Communications of the ACM* 55, no. 10 (2012): 83.

precipitating an increase in false positives. This leads to three complications. First, the total number of false and true positives cannot exceed the number of neutralizing weapons we possess.[13] Second, no matter how liberal the AI becomes in identifying targets, at no point will we be fully certain of having identified all true targets. Third, without the ability to distinguish between false and true positives, all targets will need to be neutralized. Thus, not only does this “shotgun” approach lead to intolerable social, political, and environmental costs, but it does not reach the standard of target precision required for counterforce.

Poor Representation

An image is a poor stand-in for what really differentiates vehicles: their function or role. The primary role of a vehicle’s structure is to *support* its function rather than *inform about* its function. As persons, we may employ concept learning to induce a vehicle’s role from its exterior characteristics, but AI is very much in its infancy with respect to concept learning. It is much more successful in domains where the primary purpose of an object’s structure is to inform, most notably in handwriting recognition, natural-language processing, and speech recognition. Because alphanumeric characters possess well-known properties, they are visually differentiable, in contrast to vehicles. The methods largely responsible for the AI re-awakening—convolutional neural networks (CNNs) and deep neural networks (DNNs)[14]—were designed to efficiently exploit data “regularity.”[15] Because the exterior of a vehicle provides limited information as to function, we cannot expect the same level of results in detecting TELs.

The Curse of Dimensionality

Another complication is that a two-dimensional, pixelated image is a poor representation of a three-dimensional object. From certain angles, a TEL may exactly resemble water and oil tankers or other tractor trailers. Rail-based launchers may look like regular train cars. Increasing pixel resolution and taking several images of a vehicle at different angles (possibly to form a three-dimensional model) would give better accuracy, but at high cost and diminishing returns. Increasing pixel resolution leads to the curse of dimensionality—the phenomenon by which correct AI learning “becomes exponentially harder as the… number of features… of the examples grows.”[16]

The curse of dimensionality dictates that increasing the resolution of an image from, say, 28 x 28 to 32 x 32 will increase the size of each data sample while requiring exponentially more

[13] We can push the AI to have a more liberal view of what constitutes a TEL through additional synthetic data, but only to the point where the number of targets identified is less than our ability to neutralize them.

[14] DNNs refer to neural networks with many layers of neurons. CNNs refer to neural networks in which the neurons process their inputs by combining them in ways that will show local patterns in the data.

[15] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural Computation*, Vol. 18, No. 7 (2006): 1527-1554; Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, pp. 1097-1105. 2012; and LeCun, Yann. "Generalization and Network Design Strategies." *Connectionism in Perspective* (1989): 143-155.

[16] Domingos, pp. 78-87.

samples.[17] The required exponential growth in memory will be accompanied by an exponential growth in running time for the AI.

The second consequence of the curse of dimensionality is more subtle. As resolution increases, the AI may initially detect clear patterns, but this performance will be overtaken eventually by a tendency for the images to look identical to the AI.[18] That is, images of similar objects will become dissimilar to the AI, while images of different and unrelated objects will grow more similar. We lack efficient ways for the AI to distinguish images with very high resolution. This runs counter to our own concept of thinking, by which we keep important information and discard the irrelevant. The mechanics of this process are not completely understood, and we are currently unable to replicate it with AI, given the limits of mathematical theory.

Breakthroughs such as decision trees, support vector machines, and deep learning have introduced novel, flexible discriminators. However, they all create a separating hyperplane, an idea based on Gauss's least squares. With this method, objects that are similar are closer according to some metric; objects that are dissimilar are farther apart; and a boundary separates the different objects. In high dimensions, clear separations become harder, because all data points gravitate towards the periphery of the hyper-dimensional cube.[19]

Limits to Counterforce Performance Improvement

These data limitations reveal a gap between reality and the information we receive through pixels. The question arises whether a high-powered AI can compensate for data deficiencies—can a clever AI identify clear patterns, even with imperfect data. To a degree, yes, but not well enough for counterforce standards. Five obstacles block our solving of the counterforce challenge through AI: irreducible error, no free lunch (NFL), adversarial exploitability, lack of a priori knowledge, and lack of feedback.

Irreducible Error

AI capability is limited by the quality of its data. Because no data can fully describe real-world problems, every problem carries irreducible error, based on incomplete and imperfect measurements.[20] To maximize performance given this limitation, we must decide how much we want the AI to learn from the data. How many data patterns should the AI incorporate into its decision making? The knee-jerk response may be "all of them." But some patterns are due to random events or noise, while others occur only in the data the AI is trained on. If the AI simply "memorizes" the data in the training set, it may catastrophically fail when seeing new data.[21]

[17] Ibid. p. 82.

[18] Ibid. p. 82.

[19] Ibid. p. 82.

[20] For problems in which we have complete and consistent information, there may be no irreducible error. However, these rarely occur in real-world settings, much less in the global-security domain.

[21] This is the concept behind the saying "Torture the data enough and it will confess." If a very complex learner is trained on data, it will learn patterns that do not exist, within the noise. Hence, the confession will be nonsense at best and lies at worst.

On the other hand, if we design the AI to learn only a few strong patterns, it may perform poorly—though not catastrophically poorly—on any image it sees.[22] Unfortunately, the counterforce problem leaves little room for error in striking this balance, which is referred to as the bias/variance tradeoff.

No Free Lunch

NFL is a simple concept with vexing implications. Consider the tools in an auto mechanic's garage. Each is designed for high performance in a specific, yet limited range of applications. For example, needle-nosed pliers work well for extracting objects from small crevices. But the narrow pincers cannot apply torque when unscrewing large bolts. Specialization requires tradeoffs; there is no one tool that is best for any given task. Likewise, there is not one AI algorithm that can outperform all others, solving all possible problem sets. [23] For example, we cannot say that neural networks are universally better at learning tasks than decision trees. There are no free lunches in ML.[24]

The implications of NFL significantly restrict AI performance. Because no one method is universally best, we will have to experiment with different models to find the method most suitable for the counterforce problem. Since there is an infinite number of models,[25] there can be no systematic approach to finding a model that perfectly learns the data. Even if a model were found that performs perfectly on the data we have, we cannot guarantee perfect performance on images it will see later. Invariably, we will have trained an AI that exhibits some error.

Adversarial Exploitability

There are indirect consequences to NFL. Since no AI is universally superior, even the perfect AI for identifying mobile weapons will fail if the adversary conceals, camouflages, or changes the appearance of its TELs and rail-based units, or poisons the data that leaks out. In fact, a high degree of accuracy exposes our AI to catastrophic failure. To obtain high accuracy, the AI must find those complex patterns highly correlated with known vehicle types. When the appearance of mobile weapons is changed, the complex patterns learned by the AI become irrelevant, and weapons go undetected.[26] The burgeoning research on adversarial AI has demonstrated how susceptible state-of-the-art methods are to being fooled. [27] Furthermore, we can count on

[22] This bias/variance tradeoff can be addressed somewhat by ensembles. See Domingos, p. 85. However, while variance decreases, bias is somewhat increased. Given our low margin for error, an increase in bias would be intolerable.

[23] Wolpert, David H. "The Lack of A Priori Distinctions Between Learning Algorithms." *Neural Computation*, Vo. 8, No. 7 (1996): 1341–1390.

[24] Ibid.

[25] Many methods have infinite variations.

[26] This harkens back to the bias/variance tradeoff. Because the AI is so specific to what it has seen, an adversary simply has to change up its vehicle models to fool the AI. A "dumber" AI with higher bias will perform more consistently in the face of changes, albeit at consistently lower accuracy.

[27] Three seminal works are Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing Properties of Neural Networks." *arXiv preprint arXiv:1312.6199* (2013); Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples (2014)." *arXiv preprint arXiv:1412.6572*.; and

adversaries to apply the denial and deception practices common to virtually all military forces, especially for strategic systems like nuclear-weapon delivery platforms—further straining any increase in confidence in counterforce capabilities.

The evolving nature of an adversarial environment greatly reduces the amount of knowledge we can use to tailor our AI to recognize weaponry and may render such tailoring a liability if it is based on deceptive information intentionally leaked. Unless we can account for all the ways an adversary may conceal, camouflage, or alter its mobile arsenals, we must resort to a very general AI model to avoid catastrophic error, a model whose performance falls well below the counterforce standard.

Lack of A Priori Knowledge

AI methods are much more effective when tailored to a specific problem. As LeCun states in introducing CNNs, "It is generally accepted that good [accuracy] on real-world problems cannot be achieved unless some form of a priori knowledge about the task is built into the system."[28] The more we know about the nature of the objects we are trying to detect from images, the more we can design our AI around this knowledge and obtain higher accuracy. For example, the convolutional layers of the CNNs used for handwriting recognition are designed to recognize the curves, lines, and even ink spots common to documents in a particular language. A priori knowledge in the counterforce problem is not so easily obtainable, clean, and concise. We are limited in the amount of reliable a priori knowledge we can obtain for two reasons: first, the adversary will constantly conceal, camouflage, and change the design of its mobile launchers and poison data; second, because form does not fully reveal purpose, we are limited in tailoring an AI to perceive vehicle type based on images. Because function cannot be fully represented by an image, it cannot be fully learned by AI.

Lack of Validation

The one-shot nature of a counterforce strike means that before launch there is no way to validate that the AI will work.[29] The full deployment of kinetic weapon systems is preceded by rigorous testing, including live deployment at testing sites and introduction into actual combat situations on a small scale. This process exposes many malfunctions undetectable in the course of development or the laboratory. With feedback, the weapon is fine-tuned until malfunctions virtually cease. Real-world deployment validates that a system functions as designed.[30]

Papernot, Nicolas, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. "The Limitations of Deep Learning in Adversarial Settings." In *Security and Privacy (EuroS&P),* 2016 IEEE European Symposium on Security and Privacy, pp. 372–387. For an approachable summary, see Goodfellow, Ian, Patrick McDaniel, and Nicolas Papernot. "Making Machine Learning Robust Against Adversarial Inputs." *Communications of the ACM,* Vol. 61, No. 7 (2018): 56–66.

[28] LeCun, p. 143.

[29] NFL theorems go even deeper. NFL implies that having perfect accuracy while training our AI to detect TELs does not guarantee perfect performance when going live.

[30] There are instances when the debut of a weapons systems occurs in a large-scale conflict in which it appears to perform practically flawlessly. For example, General Norman Schwarzkopf stated that the cutting-edge equipment introduced in the first Gulf War performed "beyond [their] wildest expectations." See Schwarzkopf, Norman. *It Doesn't Take a Hero: The Autobiography of General Norman Schwarzkopf*. Bantam, 2010.pp. 582–583. However, not all systems worked to perfection in

ML systems require the same process. Testing a counterforce AI with data we possess is not the same as deploying it against an adversary's nuclear arsenal. The data processed in live action may be different. Recent research in adversarial AI reveals disturbing blind spots, characterized by a brittle, narrow understanding of environments.[31] Illustrations of the types of images that fool state-of-the-art methods easily expose the limits of AI understanding[32] and the risk of catastrophic failure should a blind spot be exposed in an unanticipated circumstance. Only the live deployment of an AI can provide the quality of feedback necessary to establish its effectiveness. Unfortunately, for counterforce AI this will mean the presence or absence of a retaliatory strike.

Faulty AI Paradigms

Given these limitations, why is AI so appealing among some deterrence analysts and policymakers? It is not for lack of insight, experience, or acumen—on the contrary, their abilities and foresight tend to draw them to AI. Having successfully embraced technological changes in the past, these experts seek to do the same with this new tool. AI, however, is different from any other technology or resource we have exploited before. Past technologies derived their power principally from, and were limited by, the physical world—properties that are familiar and intuitive. AI lives principally in the computational world, which is very different from the physical and non-intuitive. Unfamiliarity with the computational world encourages three false paradigms regarding AI.

Paradigm 1: AI Solutions as Scalable

AI, and computation in general, deals with the notion that as problems become more difficult, the resources required to solve them grow exponentially. Intuition breaks down when attempting to comprehend the intractability of some problems, because sustained exponential growth in natural or human phenomena is not perceptible;[33] thus we have little context for it. When thinking about how long it will take to scale up a solution to solve a more complex problem, given the current rate of hardware and software improvement, we assume the familiar linear rate of growth. But exponential growth in the demands on resources is inevitable and quickly insurmountable. For example, solving the travelling-salesman problem (TSP)[34] for 10 cities can be done on virtually any laptop in a matter of seconds. Solving the TSP for 120 cities requires a supercomputer with as many processors as there are atoms in the universe -

that conflict. For example, a program for using satellite imagery and special forces to target SCUD missile launchers resulted in no confirmed kills. See Gordon, Michael R., and Bernard E. Trainor. *Cobra II: The Inside Story of the Invasion and Occupation of Iraq*. Pantheon, 2006, p. 179.

[31] Goodfellow, et al., "Explaining and harnessing adversarial examples."

[32] See the following two links: http://goo.gl/huaGPb, http://www.evolvingai.org/fooling.

[33] Dasgupta, S., Papadimitriou, C., Vazirani, U., *Algorithms*, McGraw-Hill, 2006, pp. 233–234.

[34] This is a classic problem where a salesman must visit n cities exactly once. Solving the problem requires finding a sequence of visits that results in the least total mileage. This generic problem has many real-world applications.

each of which must test a trillion routes per second. This supercomputer would have to run longer than the age of the universe to solve the problem.[35]

There are ways to scale up AI solutions, but they require tradeoffs. Generally, an efficient solution comes at the expense of accuracy. In the case of the TSP, an efficient solution may be obtained in minutes, but it would not necessarily be best. Just as medicines cure ailments but produce side effects, scaling up AI solutions trades one set of problems for another.

Paradigm 2: Learning Patterns is Sufficient

Armstrong et al. state that "high-level reasoning requires little computation, but low-level sensorimotor skills require enormous computational resources."[36] That is, the former—human-level intelligence—is much different from the latter—machine intelligence. AI may be trained to where it exceeds human performance in a narrowly defined problem, but as stated by Goodfellow et al., "classifiers based on modern ML techniques, even those that obtain excellent performance, … are not learning the true underlying concepts that determine the correct [classification]."[37] Through a capability called "concept learning" (whose mechanics are not completely understood), when solving problems we learn high-level concepts with relatively little data that are applicable to other problems.[38] This gives our knowledge a breadth that AI does not possess.[39] AI is not accompanied by the safeguards of common sense.

Unfortunately, intuition fails us in not making this distinction. When seeing an impressive AI result, we associate the AI behavior with human-level intelligence and an elegant, rich concept of the domain in question that it does not possess. We embrace the upside of the AI while not perceiving its restrictions. Specifically, success at one task is not necessarily transferable to another task. There is, in fact, a high likelihood of catastrophic failure if the AI is exposed to new environments. The recent surge in adversarial-AI research highlights the unstable nature of state-of-the-art ML methods. For example, instead of CNNs drawing intuitive decision boundaries among dissimilar images, almost-indiscernible changes in images can lead to misclassification.[40]

[35] Jarvis, Tyler J. "That's how the light gets in." *BYU Magazine*, Fall (2013): p. 24.

[36] Armstrong, Stuart, Kaj Sotala, and Seán S. ÓhÉigeartaigh. "The errors, insights and lessons of famous AI predictions–and what they mean for the future." *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 26, No. 3 (2014): 331.

[37] Goodfellow, et al., "Explaining and harnessing adversarial examples", p. 2.

[38] Lake, et al., "Building machines that learn and think like people", pp. 3-9.

[39] Recently, researchers developed AI that possessed a level of concept-based learning. The methods are currently confined to simpler domains. Refer to Lake, et al., "Human-level concept learning through probabilistic program induction".

[40] As stated elegantly by Goodfellow, et al., "These algorithms have built a Potemkin village that works well on naturally occurring data, but is exposed as a fake when one visits points in space that do not have high probability in the data distribution. This is particularly disappointing because a popular approach in computer vision is to use convolutional network features as a space where Euclidean distance approximates perceptual distance. This resemblance is clearly flawed if images that have an immeasurably small perceptual distance correspond to completely different classes in the network's representation." From "Explaining and Harnessing Adversarial Examples"

Paradigm 3: AI as Magic

Unlike the life sciences where there is some mystery as to how biological systems live, the mechanics of AI are fully known. As Pedro Domingos states, “Machine learning is not magic; it cannot get something from nothing.”[41] AI’s autonomous reasoning is limited to induction, a form of inference that is inferior to human conjecture and criticism.[42] It first requires data—a certain kind of data—for effective predictions. The data must be labeled, each datum associated with a classification. This data is generally abundant in the digital world, e.g., clicked-through rates, online purchases, number of likes on a post, but is not ready made for real-world problems.

Second, AI is limited as to what it can infer from data, owing to the relatively few number of higher-level mathematical concepts on which computational-learning theory is built.[43] These concepts allow us to squeeze only so much information out of data.

These paradigms are not new, nor is overestimation of AI. Generally, AI breakthroughs on a well-defined and specific problem, or narrow AI, will lead to speculation on impending artificial general intelligence (AGI). Armstrong et al. reference a study that surveyed five years of papers that contained a prediction as to when human-level artificial intelligence would be reached. There was a strong tendency among the papers, regardless of the year of publication, to predict such intelligence as 15–25 years away.[44] Lost in the hype is the incongruency between more-general and more-complicated problems.

Faulty Paradigms in the Counterforce Problem

Recent talk of AI as a potential solution to the counterforce problem has increased with advances in CNNs. In the late 1980s, Yann LeCun introduced CNNs as an improvement over fully connected neural networks because CNNs could incorporate human expertise into their design. This promise came with a warning that CNNs would be most suitable for domains that have clear, static patterns. In LeCun’s words, “In the general case specifying such knowledge may be difficult, [but] it appears feasible on some highly regular tasks such as image and speech recognition.”[45] Geoffrey Hinton introduced an efficient implementation of DL and CNNs in the

[41] Domingos, p. 81.

[42] Deutsch. “The prevailing misconception is that by assuming that ‘the future will be like the past’, it can ‘derive’ (or ‘extrapolate’ or ‘generalize’) theories from repeated experiences by an alleged process called ‘induction’. But that is impossible.” Deutsch quotes philosopher Karl Popper as saying, “We do not discover new facts or new effects by copying them, or by inferring them inductively from observation, or by any other method of instruction by the environment.”

[43] A short list of the major concepts includes the Turing machine (computability), singular-value decomposition, least squares, fast Fourier transform, central-limit theorem, and gradient descent. Most concepts used in AI are based on these concepts, a variation of them, or an auxiliary support to them.

[44] Armstrong, et al., p. 12.

[45] LeCun, p. 144.

mid 2000s[46] and, as predicted by LeCun, by 2012 breakthroughs had occurred in image[47] and speech recognition.[48] It appeared that CNNs magically found data patterns that were imperceptible to humans. Although the tasks mastered by these neural networks were simpler than the counterforce problem, it was assumed that their success could be scaled to much more difficult problems. For example, the success of Google researchers in incorporating DL into AlphaGo[49] and beating world-class Go masters led to speculation that a similar program could regulate human systems. [50]

Juxtaposed with these landmark events, however, was research in adversarial AI suggesting blind spots in DL that make them susceptible to tampering by an adversary—a particularly damning finding for the counterforce problem. Furthermore, no level of testing, data generation, or data verification could fully eliminate the blind spots.[51] This may seem odd, because seeing just a few images of a particular vehicle at various angles allows us to distinguish it from other vehicles. Our minds form rich concepts of complex objects with relatively little data.[52] AI can do this with only very simple objects at this time,[53] and there is no clear path for scaling up to more complex objects.[54] CNNs were not learning core concepts from data and, thus, had no common sense with respect to the problem they were evaluating.[55] Partly covering the blind spots ultimately requires the generation of synthetic data, exhaustive testing, and experimentation with different neural-network architectures. Thus, while DL solves one problem for a counterforce AI (feature engineering),[56] another problem takes its place (covering the blind spots).

Improving Capabilities Through a Hybrid Approach

In May 2017, RAND held a workshop that explored the possibility of AI's enabling a counterforce strike. Among distinguished experts, two views emerged. A group consisting mainly of nuclear-security experts believed AI could enable counterforce. Another group,

[46] Hinton, et al., "A Fast Learning Algorithm for Deep Belief Nets."

[47] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*. 2012: 1–9.

[48] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine*, Vol. 29, No. 6 (2012): 82–97.

[49] Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature*, Vol. 529, No. 7587 (2016): 484.

[50] Wang, Fei-Yue, et al. "Where does AlphaGo Go: From Church–Turing Thesis to AlphaGo Thesis And Beyond." *IEEE/CAA Journal of Automatica Sinica* 3.2 (2016): 113–120.

[51] Goodfellow, Ian, Patrick McDaniel, and Nicolas Papernot. "Making machine Learning Robust Against Adversarial Inputs." *Communications of the ACM* 61.7 (2018): 63–65.

[52] This comes from the ability to associate information with other, even seemingly unrelated, sources. It is not clear how our minds do that at this time.

[53] Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-Level Concept Learning Through Probabilistic Program Induction." *Science* 350.6266 (2015): 1332–1338.

[54] Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40 (2017), p. 1. Specifically, Lake et al. state, "Truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn and how they learn it."

[55] Goodfellow, et al. "Explaining and Harnessing Adversarial Examples," p.2.

[56] LeCun, Yann, Yoshua Bengio*,* Geoffrey Hinton, "Deep Learning," *Nature,* Vol. 521 (2015): 436.

composed of participants with a more technical AI background, thought object-recognizing AI would be limited, because an adversary could manipulate images to fool it.[57]

Our response is that AI will not lead to reliable counterforce, but not because of adversarial manipulation. Limits in the quality and quantity of data, coupled with the inherent limitations of ML algorithms in an evolving, deceptive domain, will prevent AI from obtaining the nearly perfect performance required for counterforce.

We emphasize that the high standard of near-perfect identification of an adversary's arsenal precludes reliance on AI. However, were we to focus on the more reachable goal of increasing the transparency of the arsenal, we could exploit Moravec's paradox ("Everything easy for a human is hard for a computer and everything hard for a computer is easy for a human") to develop a division of labor in which human analysts and computational units each do what is easy for them and hard for the other. A careful integration of the human and machine systems allows for a maximal exploitation of available data.[58] The result would not even approach what is required for counterforce, but would exceed what human analysts or AI could achieve alone.

[57] Geist, Edward and Andrew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?"

[58] For a superb example of how human and machine systems can be interleaved, see Hope, Bradley. "Inside a Quant 'Alpha Factory' --- Igor Tulchinsky's Company is Part of a Renaissance in Quantitative Investing." *Wall Street Journal*, Apr 07, 2017, *Business Premium Collection*; *Global Newsstream*.

CGSR

Center for Global Security Research
Lawrence Livermore National Laboratory
P.O. Box 808, L-189 Livermore, California 94551
https://CGSR.llnl.gov

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-TR-778038