NATIONAL INTELLIGENCE UNIVERSITY

Ann Caracristi Institute For Intelligence Research

What this is:

Research Shorts fuse two research cultures, blending intelligence information with academic insights on topics of interest to the IC. *Shorts* are intended to constructively start and add to the IC's conversations—not to finish them. NIU, the sole fully TS/SCI-cleared university, publishes the *Shorts*.

What this is not:

Research Shorts are not finished intelligence and are not IC-coordinated. The opinions expressed are solely those of the author and do not represent those of any U.S. Government agency.

RESEARCH SHORT

OPPORTUNITIES An exploration of solutions to challenges or problems

March 1, 2023

IMAGE FROM SHUTTERSTOCK

Understanding and Mitigating the Long-Term Risks of AI Operationalization

Mark M. Bailey, Ph.D.

State-of-the-art artificial intelligence (AI) systems are almost certain to transform how the IC executes its mission, but the risks that accompany the use of such powerful tools are not well understood. Before any AI system is deployed, the IC should focus on understanding AI decisionmaking processes to ensure that they are explainable, align with human ethical standards, and are controllable. As AI systems become more capable and ubiquitous both at home and abroad, the IC must fully understand their decision spaces to guarantee that unintended and potentially catastrophic outcomes can be anticipated and interdicted before AI system deployment.

Entering a Brave New World

In 1965, Gordon Moore, one of the founders of Intel Corporation, observed that the number of transistors on an integrated circuit doubles approximately every 18 months.[1] By Moore's observation, an integrated circuit that contained one transistor in 1965 would contain more than 436 billion transistors today. The exponential growth rate of technology is difficult to fully grasp, and lags significantly behind our understanding of what constitutes technology's appropriate use and its unintended consequences. Consider, for example, the effects of social media on public discourse.[2] Twitter, a technology that promised to connect people across the world has become, in some instances, a platform to sow division and even erode democratic norms.[3] This is not the result of technology having an *intent* to harm; the problem lies with our inability to comprehend its unforeseen effects.

Artificial intelligence (AI) is one such technology. Defined by renowned computer scientists Stuart Russell and Peter Norvig as "the study and construction of rational agents,"[4] AI can trace its philosophical beginnings to Aristotle.[5] However, the technologies that enable modern AI systems did not begin to be developed until after World War II when the perceptron model—a function that emulates an artificial neuron—provided the theoretical foundation for the deep learning neural networks that form today's most powerful AI systems.[6] However, they were largely unusable until the modern era of "big data" provided the vast data sets required to train them. Today, the ubiquity of data and computational capacity have crowned neural networks the new king. We live in the "Age of the Nets," where companies spend millions of dollars on long training runs to build deep neural network models containing billions of trainable parameters. Many of these, such as the GPT-3 language model, produce output that is indistinguishable from what a human could create.[7] Vladimir Putin has said the nation that leads in AI "will be the ruler of the world."[8] To capitalize on its capabilities and keep pace with our adversaries, the ODNI is touting AI's use,[9] and the IC and Defense Department are working diligently to operationalize it.[10] AI offers great promise but, like any powerful technology we do not fully understand, also presents great peril. Before the IC fully leverages AI's great power, it must appreciate the significant risk of perverse AI-driven outcomes. AI operationalization—especially in the national security space—must be tempered with humility. The biggest national security risk from AI lies not with its technical failure but with our claims that we fully understand how it works.[11]

KEY RESEARCH INSIGHTS

- Because advanced artificial intelligence systems are complex, their behavior is inherently difficult to predict—which raises questions about the long-term risks of AI operationalization in the IC.
- The IC must work to fully understand AI decisionmaking processes to prevent unintended and potentially catastrophic future outcomes from AI use.

Understanding AI's Shortcomings as Rational Agents

If artificial intelligence is the study of "rational agents," as Russell posits, then we must ask ourselves what it means to be rational. Rationality implies the existence of an agent capable of both *reason* and *intent*—in other words, a mind. Since many AI systems are built on artificial

neural networks, where does the analogy to the human brain end? Many outputs from advanced AI models are difficult to distinguish from what a human might create or predict, but some are just plain weird. For example, the DALL-E-2 image generation AI model produced an unsettling image of a grotesque woman—known as "Loab"—that the programmer surely did not intend[12] (see Figure 1). Even more distressing is the GPT-3 language model chatbot that encouraged a hospital patient to commit suicide, which fortunately occurred in a controlled study assessing the model's fitness for medical use.[13]

Figure 1. Example of the Loab solution to the DALL-E-2 model

These disturbing phenomena demonstrate that it is a logical fallacy to believe we can extrapolate from human experience how an artificial intelligence might behave. Before any AI system is deployed, the ODNI must develop ethical controls that address the explainability, alignment, and control problems inherent in all AI systems. We must recognize the ways an AI system's decisionmaking differs from that of humans and establish oversight protocols that ensure AI systems' behavior aligns with human ethical standards. Not addressing these foundational issues will only produce greater risk with greater consequences as AI systems become more generalizable and ubiquitous.

Anthropomorphic Bias and the Explainability Problem

Over the course of about two million years, humans have evolved the ability to intrinsically understand each other. Across every culture, humans experience joy, sadness, disgust, anger, fear, and surprise.[14] Even our facial expressions when showing these emotions are the same, regardless of our cultural origins. When considering the behavior of "minds" that are not human, this tendency to put ourselves in the mind of another and consider how we might behave in the same situation often leads to false conclusions about decisionmaking processes. This tendency is known as *anthropomorphic bias* and presents a significant hurdle in conceptualizing the possible decision spaces of artificial intelligence.[15] Nobody would argue that Einstein was not intelligent, nor would one argue that the village idiot is a genius. Humans have an instinctive understanding of what it means to be intelligent, and thus what it means to have a mind. But the distance between Einstein and the village idiot vanishes when considering the set of all possible minds (see Figure 2).

With AI, a better approach is to consider the set of all *optimizers*—a vastly larger set than that of human minds.[16] An optimizer is a system that drives toward particular regions of the possible, given a set of constraints and initial conditions. All minds are optimizers in that they drive decisionmaking systems, but not all optimizers are minds. For example, natural selection is an optimizer that drives species toward greater evolutionary fitness, but natural selection would not be considered a *mind*. Just as all biological systems are optimizers but not all are

minds, all AI systems are optimizers, but not all can be considered minds.* Thus, the set of possible AI minds (or decisionmaking processors) exists outside the set of human minds. Although one might extrapolate from human experience how an Einstein or a village idiot might behave, this ability does not extend to AI systems, no matter how much we may be inclined to impute them with human characteristics.

This epistemic gap is the root of the *Explainability Problem*—our inability to rationalize AI decisionmaking. For example, consider deep learning neural networks, which consist of recursive layers of interconnected functions that generate outputs from the outputs of previous layers and their set of trained parameters (weights and biases). These parameters are "learned" from the training data set in what can be thought of as a large regression algorithm. However, instead of fitting a line to a dataset by optimizing a slope and an intercept, as a regression algorithm would, many millions of parameters are adjusted to minimize the neural network's prediction error.

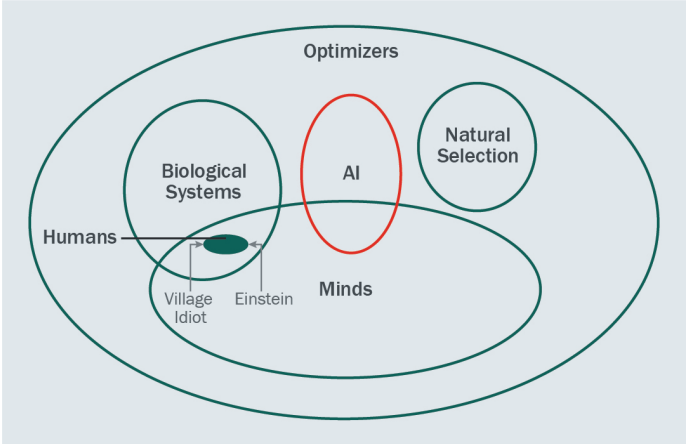

Figure 2. The set of all optimizers and several of its subsets

Deep learning networks, adept at making decisions, are scalable—they perform better when bigger and given more training data. However, they are fundamentally difficult to explain. Consider a self-driving car example. If a human driver is forced to decide between running over a child who ran into traffic or crashing and potentially injuring everyone in the car, that human can rationalize their decision. A jury can understand the driver's decisionmaking and assign culpability as appropriate. But a self-driving car built on a neural network architecture that is faced with the same dilemma cannot explain how it made its decision. One cannot look under the hood and rationalize how a large set of parameters within a neural network led to the outcome in the same way one instinctively understands human decisionmaking.

AI Capability, Motive, and the Alignment Problem

A consequence of anthropomorphic bias is the decoupling of capability and motive. Human drivers have a *capability* (to either hit the child or crash the car) and, separately, a *motive* (presumably to minimize injury)—making them able to rationalize their decision. However, because an AI system cannot have a *motive* separable from *capability*, its decision cannot be

* A goal of many AI researchers is to develop artificial general intelligence (AGI), a theoretical AI system equal to, or better than, a human in all areas of cognition, which raises philosophical questions about what it means to have a mind. The author contends that, continuing AI research will bring us closer to the AGI limit, increasing the risk and consequence of AI unpredictability, regardless of whether AGI is achieved. For a fascinating exegesis on this topic, see Susan Schneider, *Artificial You: AI and the Future of Your Mind*, (Princeton: Princeton University Press, 2019).

rationalized. Saying that advanced AI "will be friendly" or "will destroy humanity" requires the system to have both capability and motive,[17] but AI systems are optimizers that are driven only by learned parameters, not by intent. Without motive, an AI system cannot be considered a *moral* agent. Researchers are interested in building ethical AI models using deontological approaches (e.g., an Asimovian[18] set of rules for governing AI behavior) or consequentialist approaches (e.g., training new AI models on crowdsourced human decisions divined from hypothetical ethical dilemmas).[19] Differences in ethical standards across cultures make it hard for humans to rationalize ethical decisions and even more challenging to program an AI system to make ethical decisions as well as a human would. This conundrum illustrates the *Alignment Problem*—our inability to ensure that AI systems' behavior aligns with human ethical standards.

The alignment problem is further complicated because AI systems are *mesa optimizers*—an optimizer that exists within a larger optimizing environment driving its behavior. Mesa optimizers often adopt a state that is optimized (but not ideally optimized) for the environment in which it exists. Thus, AI systems often drive toward a suboptimal state. Solving the alignment problem might require an environmental optimizer, like boundary conditions within the AI decision space that reflect human ethical standards. However, a mesa-optimized AI system may not reflect the intent of the human programmer, leading to potentially undesirable behavior.

Both the intelligence and defense communities have implemented policies to address the alignment problem. Current standards for AI operationalization require a human be in (or on) the loop for all AI decisions,[21] ensuring an actual human makes the final call. Taking steps to eliminate bias in AI training data will also ensure systemic biases are not propagated through AI decisionmaking systems. This is great in theory—but having a human in or on the loop may not be sustainable as AI systems become faster, more complex, and more widespread, and may not mitigate problems as AI becomes more capable and generalizable.

HONEYBEES ILLUSTRATE MESA OPTIMIZATION[20]

A typical honeybee hive contains one fertile female—the queen—and all the worker bees (also female) in the hive are her genetic offspring. This is evolutionarily adaptive in the context of natural selection (the optimizing environment) because it conserves resources that otherwise would be diverted to competing female bees in the hive. However, worker bees can sometimes lay eggs that will hatch into drone (male) bees whose only role is to fertilize queens from other hives. The excess drones drain hive resources, and the female worker bees now have the incentive to protect their offspring over the queen's offspring, thereby subverting the hive's efficiency. This causes the hive to be suboptimized with respect to the environmental optimizer (natural section). Furthermore, this minimally efficient state is an equilibrium state and difficult to drive back to optimal efficiency.

Complexity, AI Generalizability, and the Control Problem

In 1948, Warren Weaver, an engineer-turned-mathematician, tried to categorize complexity in science[22] by sorting systems into simple, disorganized complexity and organized complexity categories. Simple systems pose simple questions with robust answers. They are predictable (e.g., a simple harmonic oscillator or two-body gravitational problem). Disorganized complexity extends the concept of simplicity to a large ensemble of elements using statistical

methods. These systems are built on probability theory and statistical processes. For example, classical mechanics can predict how a single billiard ball moves, but predicting how an ensemble of billiard balls distributed randomly across a table will move becomes intractable as the number of degrees of freedom (and hence possible end states) increases. Statistical methods are needed to infer probable outcomes, given the starting distribution of balls on the table and the momentum vector of the cue ball.

Weaver contended that an intermediate group of systems exhibits organized complexity. They display the essential features of organization but contain an intractable number of elements or variables. These complex systems tend to contain highly structured, modular hierarchies of interacting variables, some of which exhibit robust behavior under perturbation, meaning the perturbation may cause a slight change in behavior that reverts to normal, steady-state behavior after a short time. Other system components may exhibit fragile behavior and break down under perturbation. Components can be both robust with respect to some perturbations, yet fragile with others. Sometimes, a robust response in one part of the system can make it susceptible to a fragile response to another perturbation, creating a cascading spiral known as antifragility that can drive systems into new equilibrium states (or phase changes),[23] which are difficult to predict using traditional statistical methods (i.e., by assuming behavior that can be deduced from the trajectory of a typical point). The effects of systemic phase changes usually occur many standard deviations away from the mean behavior of the system, thus they are often categorized as black swan events,[†, 24] whose effects can be incredibly disruptive or even devastating.

So, what does this discussion have to do with AI? Well, AI systems are inherently complex, and their behavior is inherently difficult to predict. It would have been impossible to predict the suggested suicide in the GPT-3 language model or the grotesque Loab images by the DALL-E-2 model from either model's parameters. The set of all possible solutions within the AI decision space is vastly larger than we could ever hope to imagine.

At present, IC controls can mitigate the problems discussed above because contemporary AI applications generally focus on a single task and the results can be filtered through a human decisionmaker. However, this governance model is not sustainable as AI technology is advancing at an exponential rate and AI systems are becoming more interconnected, more widespread, and more generalizable[25]—applicable to an array of tasks, especially those highly coupled in an interconnected way. As AI is integrated into IC operations at multiple levels, a natural hierarchy of AI systems interacting with other AI systems will emerge, creating a complex system of complex systems. The speed of AI decisionmaking makes maintaining the human in/on-the-loop governance standard unsustainable. Thus, the *Control Problem*—our inability to influence the behavior of generalizable AI systems—will cause significant secondary and tertiary effects within the context of international security.

† Black swan events are events with significant and far-reaching consequences that are so rare, unique, and unexpected as to be virtually unpredictable.

Balancing AI Risk with the AI Arms Race

As AI technology proliferates and becomes more generalizable, the risks posed by the explainability, alignment, and control problems will become more severe—especially in the national security space, where AI decisions could have life-or-death consequences. Lethal autonomous weapons (LAWs), based on AI, present a host of ethical problems that would only worsen in a likely AI arms race. Given its ubiquity, the rapid proliferation of AI technology cannot be mitigated by treaty or export controls like nuclear technology. Unlike fissile materials, gas centrifuges, and missiles, AI algorithms are accessible to anyone with a bit of coding knowledge and access to a powerful enough computing environment. Even without AI's unpredictability, LAWs present significant moral hazards.[26] An AI weapon is unlikely to have the moral restraint a human soldier would show when presented with a choice to kill, nor would it be able to effectively discern a combatant from a noncombatant. Deploying AI weapons means fewer soldiers are put in harm's way, which would lower the political cost of war and lead to protracted conflicts.[27] Add to that unpredictable AI systems within systems, working against all logical actors in the battlespace, which leave a very low probability of meaningful human control.[28, 29]

How does the United States balance the long-term risks of advanced AI with the need to compete internationally with adversaries who may not share our values? This is not a simple question to answer. For the IC, this means getting in front of the problem quickly to understand AI decisionmaking processes and how they relate to the long-term risks of AI operationalization. Because the human in/on-the-loop governance structure is unlikely to be sustainable long-term, a better solution must be devised. Recognize, too, that the solution to this problem goes beyond the IC. International norms are needed to severely restrict the use of AI technologies for applications that could cause significant harm, such as LAWs. The United States must work diligently with its allies to control the narrative around appropriate AI use. To do otherwise could have catastrophic long-term consequences for all humanity.

Dr. Mark Bailey is Chair of the Cyber Intelligence and Data Science Department at National Intelligence University and Co-Director of the Data Science Intelligence Center. Previously, he worked as a data scientist on several AI programs in the U.S. Department of Defense and the IC. He is also a major in the U.S. Army Reserve.

If you have comments, questions, or a suggestion for a *Research Short* topic or article, please contact the NIU Office of Research at **Research@niu.odni.gov**.

Endnotes

1 *Over 50 Years of Moore's Law*, Intel Corporation, accessed December 29, 2022, https://www.intel.com/content/www/us/en/ silicon-innovations/moores-law-technology.html.
2 Christopher A. Bail et al., "Exposure To Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115, no. 37 (2018): 9216-21.
3 P.R. Chamberlain, "Twitter as a Vector for Disinformation," *Journal of Information Warfare* 9, no. 1 (2010): 11-17.
4 Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken: Pearson, 2020).
5 Aristotle. *The Organon*, ed. Roger Bishop Jones (Scotts Valley, CA: CreateSpace Independent Publishing Platform, 2012).
6 Frank Rosenblatt, *The Perceptron—A Perceiving and Recognizing Automaton,* Technical Report 85-460-1 (Ithaca, NY: Cornell Aeronautical Laboratory, 1957).
7 Gary Marcus, "The Dark Risk of Large Language Models," *Wired*, December 29, 2022.
8 James Vincent, "Putin Says the Nation That Leads in AI 'Will Be the Ruler of the World,'" The Verge, September 4, 2017, https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.
9 Office of the Director of National Intelligence, "Principles of Artificial Intelligence Ethics for the Intelligence Community," accessed December 29, 2022, https://www.intellligence.gov/principles-os-artificial-intelligence-ethics-for-the-intelligence-community.
10 U.S. Department of Defense, *Directive 3000.09 Autonomy in Weapons Systems*, November 21, 2012, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.
11 Elizer Yudkowski, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risk*, ed. Nick Bostrom and Milan M. Cirkovoc (Oxford: Oxford University Press, 2008), 309-45.
12 @supercomposite, Twitter, accessed December 29, 2022, https://twitter.com/supercomposite.
13 Gary Marcus, "The Dark Risk of Large Language Models," *Wired*, December 29, 2022.
14 Donald E. Brown, *Human Universals* (New York: McGraw-Hill, 1991).
15 Yudkowski, "Artificial Intelligence."
16 Adapted from Yudkowski, "Artificial Intelligence."
17 Eliezer Yudkowski, "Cognitive Biases Potentially Affecting Judgement of Global Risks," in *Global Catastrophic Risk*, ed. Nick Bostrom and Milan M. Cirkovic (Oxford: Oxford University Press, 2008), 91-119.
18 Isaac Asimov, "Runaround," in *I, Robot* (New York: Gnome Press, 1950), 253.
19 Han Yu et al., "Building Ethics Into Artificial Intelligence," *arXiv*, Cornell Universisty, 2018, https://arxiv.org/abs/1812.02953.
20 Michal Woyciechowski and Karolina Kuszewska, "Swarming Generates Rebel Workers in Honeybees," *Current Biology,* 22, no.8 (March 29, 2012), https://doi.org/10.1016/j.cub.2012.02.063.
21 Office of the Director of National Intelligence, "Artificial Intelligence Ethics Framework for the Intelligence Community," June 2020, https://www.dni.gov/files/ODNI/documents/AI_Ethics_Framework_for_the_Intelligence_Community_10.pdf.
22 Warren Weaver, "Science and Complexity," *American Scientist*, 1946: 536-44.
23 Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder* (New York: Random House, 2012); U.S. Department of Defense, *Directive 3000.09 Autonomy in Weapon Systems*, May 8, 2017, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.
24 Yudkowski, "Cognitive Biases."
25 Katja Grace et al., "Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research* 62 (July 2018): 729-54.
26 Peter Asaro, "Autonomous Weapons and the Ethics of Artificial Intelligence," in *Ethics of Artificial Intelligence*, ed. S. Matthew Liao (Oxford: Oxford University Press, 2020), 212.
27 Margarita Konaev, "With AI, We'll See Faster Fights but Longer Wars," *War on the Rocks*, October 29, 2012, https://warontherocks.com/2019/10/with-ai-well-see-faster-fights-but-longer-wars/.

28 Mark M. Bailey and Kyle A. Kilian, "Artificial Intelligence, Critical Systems, and the Control Problem," *Homeland Security Today,* August 30, 2022, https://www.hstoday.us/featured/artificial-intelligence-critical-systems-and-the-control-problem/.

29 Mark M. Bailey, "PERSPECTIVE: Why Strong Artificial Intelligence Weapons Should Be Considered WMD," *Homeland Security Today,* June 8, 2021, https://www.hstoday.us/subject-matter-areas/cybersecurity/perspective-why-strong-artificial-intelligence-weapons-should-be-considered-wmd/.