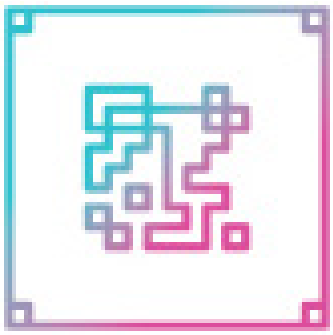
JUNE 2018



ARTIFICIAL INTELLIGENCE

What Every Policymaker Needs to Know

Paul Scharre and Michael C. Horowitz
Preface by Robert O. Work

Center for a New American Security

About the Authors

MICHAEL C. HOROWITZ
is an Adjunct Senior Fellow at the Center for a New American Security (CNAS) and a professor of political science at the University of Pennsylvania.

ROBERT O. WORK
is Senior Counselor for Defense at CNAS and former Deputy Secretary of Defense.

PAUL SCHARRE
is a Senior Fellow and Director of the Technology and National Security Program at CNAS.

About the Technology & National Security Program

Technology is changing our lives. Rapid developments in artificial intelligence, autonomy, cyber-physical systems, networking and social media, and disinformation are profoundly altering the national security landscape. Nation-states have new tools at their disposal for political influence as well as new vulnerabilities to attacks. Non-state groups and individuals are empowered by social media and radical transparency. Artificial intelligence and automation raise profound questions about the role of humans in conflict and war.

CNAS' Technology and National Security program explores the policy challenges associated with these and other emerging technologies. A key focus of the program is bringing together the technology and policy communities to better understand these challenges and together develop solutions.

About this Report

This report is part of the Center for a New American Security's series on Artificial Intelligence and International Security. The series examines the potential consequences of advances in artificial intelligence for the national security community. Nearly every aspect of national security could be transformed by artificial intelligence. AI has applications for defense, intelligence, homeland security, diplomacy, surveillance, cybersecurity, information, and economic tools of statecraft. The United States must not only anticipate these developments, but act decisively to prepare for uses by competitors and take advantage of the opportunities AI presents.

Also in this Series

The Artificial Intelligence and International Security series includes:

- **Artificial Intelligence and International Security** by Michael C. Horowitz, Paul Scharre, Gregory C. Allen, Kara Frederick, Anthony Cho, and Edoardo Saravalle (forthcoming)

- **Strategic Competition in an Era of Artificial Intelligence** by Michael C. Horowitz, Elsa Kania, Gregory C. Allen, and Paul Scharre (forthcoming)

This series is part of the Center for a New American Security's multi-year Artificial Intelligence and Global Security Initiative. Learn more at cnas.org/AI.

Acknowledgements

We would like to thank Loren Schulman for her helpful comments on an early draft of this report and Jack Clark for his feedback on elements of this report. We would also like to thank Maura McCarthy, Jacki Fink, and Allene Bryant for their role in the production and design of this report. Any errors or omissions are the sole responsibility of the authors. CNAS does not take institutional positions.

Cover Photo
Tristan Campos/CNAS

ARTIFICIAL INTELLIGENCE

What Every Policymaker Needs to Know

Preface

By Robert O. Work

We are in the midst of an ever accelerating and expanding global revolution in artificial intelligence (AI) and machine learning, with enormous implications for future economic and military competitiveness. Consequently, there is perhaps no debate more important than how the United States and other democratic powers exploit advances in AI and the associated technologies, sub-disciplines, and methods used to create intelligent machine behavior – within the moral, ethical, political, and legal boundaries acceptable to their leaders and citizens.

The idea of establishing boundaries for AI is vitally important in democratic societies. The general public reaction to the prospect of a future where more and more tasks and decisions are delegated to machines is decidedly mixed, having been indelibly shaped for decades by science fiction writing, television, and movies. As early as 1927, Metropolis, one of the first full-length sci-fi movies ever made, told the story of a scientist who builds a robot to replace his lost love. But the robot gets other ideas and ultimately holds sway over an entire city. Metropolis is just one example of two big concerns about intelligent machines often explored in science fiction: They will either enslave us (e.g., Metropolis, the Matrix trilogy), or they will kill us (e.g., The Terminator, Battlestar Galactica). And now, given how AI and machine learning are beginning to impact the workplace, there is a third compelling concern: Intelligent machines will take our jobs (e.g., Martin Ford's book, Rise of the Robots: Technology and the Threat of a Jobless Future).

These dystopian outcomes need to be balanced by visions of a future in which intelligent machines have a more positive impact on our society. AI and machine learning will likely lead to a new industrial revolution, improving economic competitiveness and creating new sources of wealth. They will lead to advances in medical science and automobile safety. They will enable new forms of virtual training and entertainment. Their positive impact on our society and well-being is likely to be profound.

And they will inevitably impact international security and the application of military power – the subject of this report. The following is intended as a primer on AI and machine learning in the national security space. It explains the AI language and ideas policymakers need to know; explores the security-related applications of artificial intelligence; ponders strategic competition in an era of AI and machine learning; and discusses the indirect effects of the AI revolution for global security.

We hope the report will provide a solid foundation for a healthy debate on how AI can be used responsibly to improve our national security. Indeed, it provides the intellectual vector for the CNAS Task Force on Artificial Intelligence and National Security, which I co-chair with Dr. Andrew Moore of Carnegie Mellon University. Driven by the desire to debate thoroughly the acceptable boundaries of AI and machine learning in security applications, the task force – consisting of a variety of experts from government, academia, and public/private businesses and organizations – will discuss such topics as:

- Ensuring U.S. leadership in AI research and innovation
- Empowering the federal government to take advantage of AI opportunities
- Ensuring safe and responsible uses of AI in national security applications
- Preparing to counter the malicious uses of AI

It will also try to find the right balance between the more pessimistic and optimistic narratives associated with AI and machine learning, and ensure the ethical and moral pursuit of these technologies.

I hope you enjoy the read!

The Artificial Intelligence Revolution

The artificial intelligence revolution is underway. Tremendous gains in AI and machine learning are being applied across a range of industries: medicine, finance, transportation, and others. These developments likely will have a profound impact on the global economy and the international security environment. Business leaders and politicians around the world, from Elon Musk to Vladimir Putin, are increasingly thinking about whether AI will trigger a new industrial revolution. Like the steam engine, electricity, and the internal combustion engine, AI is an enabling technology with a wide range of applications. The technologies in the first and second industrial revolutions allowed the creation of special-purpose machines that could replace human physical labor for specific tasks. Today, AI is enabling the creation of special-purpose machines to replace human cognitive labor for specific tasks. As co-founder of Wired Kevin Kelly observes, "[AI] will enliven inert objects, much as electricity did more than a century ago. Everything that we formerly electrified we will now cognitize."[1]

Preparing for the consequences of the AI revolution is a critical task for the national security community.

Preparing for the consequences of the AI revolution is a critical task for the national security community. Nearly every aspect of national security could be shaped by artificial intelligence. AI has applications for defense, intelligence, homeland security, diplomacy, surveillance, cybersecurity, information, and economic tools of statecraft. The United States must not only anticipate these developments, but act decisively to prepare for uses by competitors and take advantage of the opportunities AI presents.

It is not enough, however, to prepare only for AI's direct applications to national security missions. The first and second industrial revolutions kicked off a broad pattern of industrialization that led to sweeping social, economic, and political change. Nations rose and fell. Urbanization and industrialization changed domestic politics and led to the rise of the middle class. Even the key metrics for global power changed, with coal- and steel-producing nations gaining in strength and oil becoming a global strategic resource. The geography of power also changed as nations fought to secure access to critical resources, culminating in wars over territories that would have been insignificant in an era of agricultural power.

The AI revolution could also change the balance of power and even the fundamental building blocks of the global economy. Just as coal fuels steam engines and oil fuels internal combustion engines, data fuels the engines of machine learning. Nations with access to the best data, computing resources, human capital, and processes of innovation are poised to leap ahead in the era of artificial intelligence. As the world's most advanced economy and an engine of technological innovation, the United States has many advantages over other nations, but it is not alone in this technology race. China is a major player in AI and has embarked on a national plan to be the world's leader by 2030. Russia has signaled its interest in AI, with Putin stating in 2017 that "the one who becomes the leader in [artificial intelligence] will be the ruler of the world."[2]

The integration of AI technologies across human society could also spark a process of cognitization analogous to the changes wrought by industrialization. Automation will transform and replace jobs, alter the balance between labor and capital, and change national politics and foreign policy. A study conducted by the McKinsey Global Institute recently estimated that roughly 45 percent of job tasks currently being done in the U.S. economy could be automated using existing technology. How societies manage these changes will affect their internal cohesion and global competitiveness.

This report is part of the Center for a New American Security's multi-year Artificial Intelligence and Global Security Initiative. It is intended as an introduction to the impact of advances in artificial intelligence for national security and an initial exploration into how AI may change the international security environment. It builds on work done by experts from CNAS and other institutions. The AI revolution will take decades to unfold and will evolve in surprising ways. Computers and digital networks have evolved considerably since the early days of mainframes and ARPANET. The new digital information landscape of social media, viral videos, and fake news would have been hard to foresee half a century ago. It is impossible to foresee all of the possible

The integration of AI technologies across human society could also spark a process of cognitization analogous to the changes wrought by industrialization.

changes that artificial intelligence and machine learning may bring to global security, but with preparation policymakers can better chart a course through the uncertain waters ahead.

What is Artificial Intelligence?

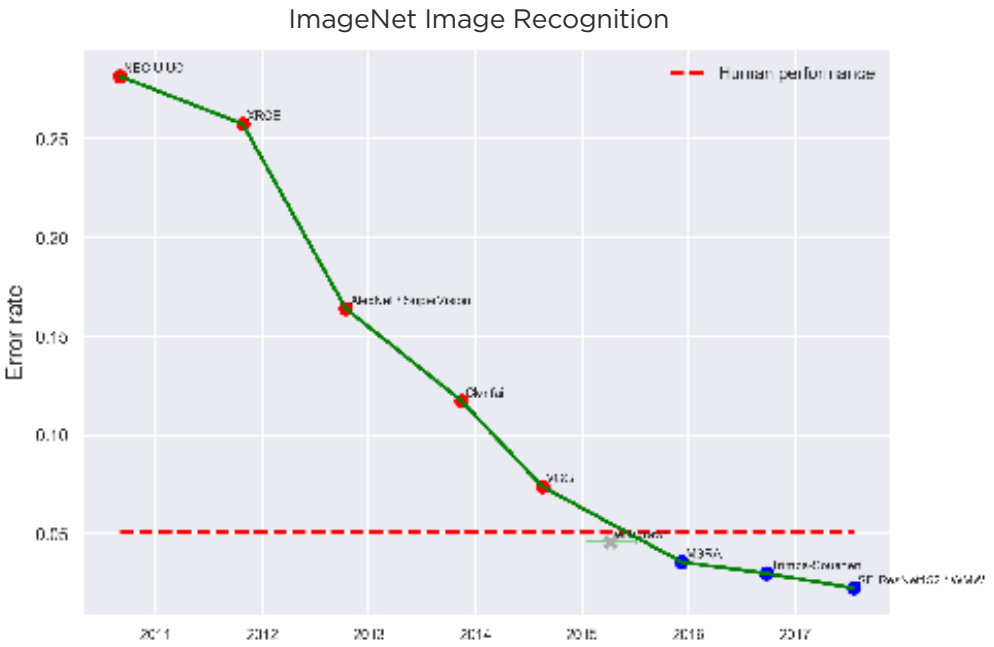
Artificial intelligence and machine learning, a method of AI, make it possible to build special-purpose machines to perform useful cognitive tasks, in some cases better than humans. Early AI systems were rule-based "expert systems" where a computer program simply followed a set of specific instructions about how to behave in a particular situation. Recent AI advances enable much more sophisticated systems. Machine learning allows algorithms to learn from data and develop solutions to problems. These increasingly intelligent machines can be used for a wide range of purposes, including analyzing data to find patterns and anomalies, predicting trends, automating tasks, and providing the "brains" for autonomous robotic systems.

Current AI systems are "narrow," however, in that their expertise is confined to a single domain, as opposed to hypothetical future "general" AI systems that could apply expertise more broadly. Machines – at least for now – lack the general-purpose reasoning that humans use to flexibly perform a range of tasks: making coffee one minute, then taking a phone call from work, then putting on a toddler's shoes and putting her in the car for school.

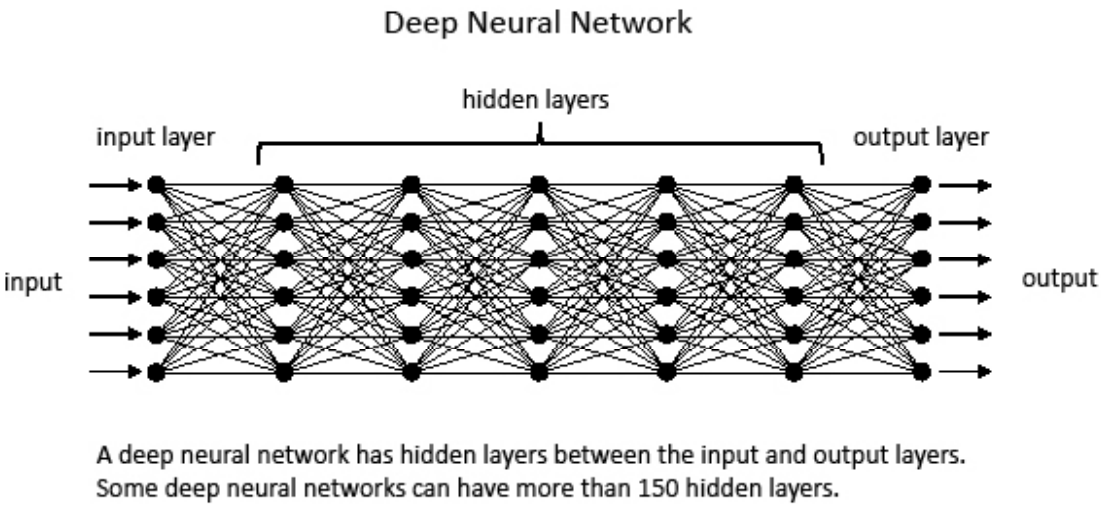
Current AI systems are "narrow," however, in that their expertise is confined to a single domain, as opposed to hypothetical future "general" AI systems that could apply expertise more broadly.

This narrowness is a significant limitation of AI systems. Current AI systems can fail if they are deployed outside of the context for which they were designed, making their performance "brittle" in real-world applications. AI systems that outstrip human abilities in one task may suddenly perform poorly if the context for their use changes. Despite these limitations, narrow AI systems have tremendous value and are already being applied to a range of real-world arenas, from stock trading to shopping to predicting the weather.

Artificial intelligence is the field of study devoted to making machines intelligent.[3] Intelligence measures a system's ability to determine the best course of action to achieve its goals in a wide range of environments.[4] The field of AI has a number of sub-disciplines and methods used to create intelligent behavior, and one of the most prominent is machine learning.



AI-based image recognition systems surpassed human performance in 2016 in benchmark tests, such as the ImageNet image database. (Electronic Frontier Foundation, AI Progress Measurement)



A deep neural network has hidden layers between the input and output layers. Some deep neural networks can have more than 150 hidden layers. (Paul Scharre)

Machine Learning

Machine learning has proven to be a particularly powerful approach for generating intelligent behavior. Given a goal, learning machines adjust their behavior to optimize their performance to achieve that goal.[5]

Data is the fuel that powers the engine of machine learning. Supervised learning makes use of labeled training data. For instance, an algorithm might take as input millions of labeled images, such as "dog," "person," "apple." The algorithm then learns subtle patterns within the images to distinguish between categories – for example, between an apple and a tomato. This approach, which relies on large amounts of data and machine learning, can be helpful in situations where a rule-based approach might come up short. Trying to hand-code a set of rules for a machine (or a person) to visually distinguish between an apple and a tomato would be challenging. Both objects are round, red, and shiny with a green stem on top. Yet they look different in subtle and important ways that are obvious even to a young child. Given enough labeled images of both, machines can also learn these differences and then distinguish between an apple and a tomato when they are not labeled. In fact, in 2016 machines surpassed humans at benchmark tests for image classification.[6]

Data is the fuel that powers the engine of machine learning.

Unsupervised learning uses unlabeled training data – like the same images of apples, tomatoes, or dogs, but with no name attached to them. Even without labels, machines can sort data into clusters or categories based on patterns within the data. Watching a new sports game, humans can discern patterns of behavior and rules, even if the terms used in the game may be foreign. Similarly, machines can find anomalies or predict future behavior by analyzing data. For example, an AI system can analyze financial transactions and sort them into clusters based on data associated with the transaction (time, amount, sender, etc.) This can be used to identify anomalous transactions that are outside the norm and may be fraudulent activity. This has wide-ranging applications, from detecting brain tumors to predicting the weather.

Reinforcement learning uses feedback from the environment to train machines.[7] Just as humans learn from touching a hot stove, AI systems can learn from environmental feedback whether their actions are helpful or harmful in accomplishing their goals. For example, AI systems have learned to play Atari games based on feedback from the game score. The AI system learns that some moves result in a higher score; over time, this can improve the system's behavior to optimize pursuit of a goal, such as winning the game. In some cases, even without human training data, AI systems using reinforcement learning have learned to play games to superhuman levels of performance.

Deep learning is a type of machine learning that uses neural networks. Neural networks are loosely inspired by biological neurons and use a series of artificial neurons connected in a layered network. Input data flows into one end of the network, then signals cascade across the network through the artificial neurons to an output layer. For example, the input data for a neural network

doing image recognition would be each pixel in an image. The output of the neural network would be the label for that image. "Deep" neural networks are those that have multiple "hidden layers" between the input and output layer. Neural networks learn by adjusting the weight of the connections between each neuron to optimize the paths through the network to achieve a certain output. Some deep neural networks used for image recognition can have hundreds of thousands of artificial neurons.[8] Neural networks can learn via supervised learning, unsupervised learning, or reinforcement learning, depending on whether the data used to train the neural network is labeled, unlabeled, or comes from environmental feedback. Deep neural networks are widely used for many AI applications today, from image recognition to predicting medical outcomes.

Generative adversarial networks pit two competing neural networks against one another in a game. One network attempts to create synthetic (computer-generated) data that is indistinguishable from training data. For example, this could consist of creating a computer-generated image of a dog, based on training data pictures of real dogs. Using the output of the first neural network, an "adversarial" network attempts to discern the synthetic data from the training data – distinguishing between the real dogs and the computer-generated ones.[9] With each iteration, both networks learn and improve. This approach has been used to generate "deep fakes" – high-quality fake pictures, audio, and video that are indistinguishable to humans from the real thing.[10]

Machine learning is one approach for building intelligent machines, but it is not used in all forms of AI. For example, the poker-playing AI system Libratus that defeated top human poker players in 2017 uses computational game theory and does not use machine learning. Other forms of AI include: neural networks, evolutionary or genetic algorithms, computational game theory, Bayesian statistics, inductive reasoning, fuzzy logic, analogical reasoning, and hand-coded expert knowledge, to name just a few.

Advances in Machine Learning Methods

Successful machine learning applications generally require large amounts of data to train algorithms.

For instance, the ImageNet database used to train image classifiers has over 14 million labeled images, and the organizers have a goal of 1,000 images per image category (for example, a dog would be an image category, as would a soccer ball).[11] The first version of AlphaGo, a program developed by AI research company DeepMind in 2016, initially learned to play the Chinese strategy game Go based on training data from 30 million human moves. However, data can be a major limitation for applications where large datasets may not exist. In such settings, AI researchers are increasingly turning to "synthetic data" created via computer simulations. A newer version, AlphaGo Zero, did not use any initial training data from human games; it learned to play Go by playing against itself. Large amounts of synthetic data were still needed, however, with AlphaGo Zero playing 4.9 million games against itself.[12]

AI researchers are increasingly turning to "synthetic data" created via computer simulations.

AI researchers are also improving their ability to train machines using sparse datasets. Google's multilingual, neural-network-based language translation tool has been able to do "zero-shot" translation between two languages for which it has no translation data by relying on data between each language and a third language. By feeding in data on Portuguese to English and English to Spanish translations, the system learned to translate Portuguese to Spanish without any Portuguese to Spanish training data.[13] This could allow translations between rare languages for which there may be little data from one to the other, but some data translating each to a more common language such as English.

A major hurdle for AI systems today is their limitations in transferring learning from one task to another related task. Humans can learn one skill, then leverage that knowledge to more quickly acquire knowledge in a related area, building on what they already know. When AI systems attempt to learn a new task, they frequently suffer from "catastrophic forgetting," where they lose their old knowledge. In December 2017, DeepMind released AlphaZero, a single algorithm that could learn to play Go, chess, or the Japanese strategy game Shogi.[14]

A major hurdle for AI systems today is their limitations in transferring learning from one task to another related task.

Building a single algorithm that could learn to play three different strategy games without any training data was an impressive feat. Different versions of AlphaZero needed to be trained for each game, however. AlphaZero could not transfer learning from one game to another, as

a human might. This limitation restricts AI systems to narrowly performing only one task, even if they acquire superhuman performance at that task.

AI researchers are making progress on multi-task learning, however. In February 2018, using deep reinforcement learning, DeepMind trained a single AI system to perform 30 different tasks within a simulated environment. Not only did the agent learn new tasks without forgetting others, the agent's performance demonstrated positive transfer of learning between some tasks.[15]

AI Progress through Games

Games have often played a critical role in the advancement of AI, both as a challenge for researchers and as a benchmark for progress. The first game to fall to machines was tic-tac-toe (noughts and crosses), beaten in 1952. Chess was an early target, with programmers building the first chess-playing computers in the 1950s, but these early programs were far short of human abilities.[16] Checkers, which is simpler than chess, fell to machines in 1994. A few years later, in 1997, IBM's DeepBlue beat Gary Kasparov at chess.

Chess, checkers, and Go provide useful yardsticks for AI progress because their complexity can be quantified. For example, the total space of possible positions in checkers is 5 x 10^20 (500 billion billion possible positions).[17] In 2007 AI researchers "solved" checkers by calculating the optimal move for every relevant position (roughly 10^14 positions). By "solving" checkers, AI researchers were able to do far more than simply beat human performance; they were able to determine the best move in any given situation.

Chess, checkers, and Go provide useful yardsticks for AI progress because their complexity can be quantified.

Chess is far more complex, with roughly 10^40 to 10^50 possible positions. This means that heuristics (common rules for behavior) are needed to win at chess, which cannot be computationally solved. Go is another matter entirely, with approximately 10^170 possible positions. This is roughly 10^100 (a googol) more complex than chess and more than the number of atoms in the known universe.[18] For Go, the number of calculations to mathematically solve the game is so large that the same kinds of brute force methods used in checkers and early chess programs are inadequate.

Over time, AI researchers have taken to tackling more open-ended games in a variety of areas. The quiz show Jeopardy!, for example, has a much more unbounded space of potential questions than strategy board games and requires reasoning by analogy, understanding puns and riddles, and other linguistic challenges. IBM's Watson defeated human contestants Ken Jennings and Brad Rutter at Jeopardy! in 2011, in part due to its superior reflexes at timing when to buzz in.

Jeopardy! contestants Ken Jennings and Brad Rutter lost to IBM's supercomputer 'Watson' in Jeopardy! in 2011. Defeating humans at quiz games involving puns, analogies, and other linguistic challenges was a major step forward for AI. (Ben Hider/Getty Images)

Google DeepMind's AI program AlphaGo plays against Lee Sedol in 2016. AlphaGo was able to reach superhuman levels of play by playing against itself. (Kim Min-Hee-Pool/Getty Images)

Poker has long been seen as particularly challenging for AI systems since it is an incomplete information game. Unlike checkers, chess, or Go, where all of the relevant information is on the board, in poker some information (the other players' cards) is hidden. This is a much more complex challenge, but the AI Libratus defeated top human poker players in 2017. Interestingly, Libratus was able to do so without using many of the techniques that human players use, such as spotting tells or capitalizing on other players' weaknesses.[19]

AI researchers have also tackled computer games of increasing complexity. In 2014, DeepMind developed a single algorithm that learned to play a range of different Atari games with only the pixels from the screen and game score as input. Using reinforcement learning, the algorithm was able to learn which moves improved its score. Their algorithm was able to play near or above human level at more than half of the 49 games it played. (A different algorithm had to be trained for each game, due to catastrophic forgetting; the moves in Pac-Man are different than the moves in Asteroids, so learning how to play one does not help an AI learn to play the other.)

When AI researchers at DeepMind developed a super-human Go program, they did so using a combination of methods. First, they used data from 30 million human moves to train the algorithm how humans play, a form of supervised learning. Then, they had the machine play against itself to evolve its game even further through reinforcement learning.[20] This initial program, AlphaGo, beat top human player Lee Sedol in 2016. In November 2017, DeepMind released a new version, AlphaGo Zero, which taught itself to play entirely through self-play and without any human examples. Within three days of self-play, during which it played 4.9 million games, AlphaGo Zero achieved superhuman performance and beat the previous version of AlphaGo 100 games to zero.[21] A few weeks later in December 2017, DeepMind released AlphaZero, a single algorithm that learned to play Go, chess, and Shogi through reinforcement learning (with a different version for each game). In the case of chess, AlphaZero was able to reach superhuman performance – eclipsing millennia of human knowledge at chess – after a mere four hours of self-play.[22]

Recently, AI researchers have turned to computer realtime strategy games as a testbed for AI.

More recently, AI researchers have turned to computer real-time strategy games as a testbed for AI. In these games, such as Starcraft, players simultaneously compete in an open environment, controlling and fighting multiple units at one time, making them more computationally complex than turn-based games such as chess. The AI research company OpenAI developed a program that beat humans at the computer game Dota 2

in 1v1 play in 2017.[23] OpenAI used a similar technique as the first AlphaGo, with initial supervised learning based on human play and then self-play using reinforcement learning to reach superhuman abilities.

Looking ahead, AI research companies are focusing on ever more complex strategy games. Dota 2 is normally played in 5v5 team matches, and OpenAI researchers have announced they are turning their attention to 5v5 play. DeepMind has said they are designing an algorithm to beat humans at Starcraft, another real-time strategy game.[24]

Applicability of Current AI Methods

Progress in games demonstrates the art of the possible with current AI methods. Deep learning, and deep reinforcement learning in particular, has proven to be a powerful method for tackling many different problems.[25]

Competitive self-play is valuable in improving AI performance, from creating fake images to achieving superhuman performance at games. As the AI research company OpenAI explained in a blog post regarding their system that beat humans at Dota 2:

> [S]elf-play can catapult the performance of machine learning systems from far below human level to superhuman, given sufficient compute [computing power] ... Supervised deep learning systems can only be as good as their training datasets, but in self-play systems, the available data improves automatically as the agent gets better.[26]

For tasks that have a clear metric for better performance, a sufficiently bounded space of possible options, and training data or the ability to generate synthetic data, machine learning can sometimes yield human- or superhuman-level performance. These techniques are applicable to a wide variety of real-world tasks.

What is AI Good For?

Rule-based AI systems have been around for decades, but recent advances in big data, computational power, and improved algorithms have led to significant improvements in AI capabilities. As a result, more advanced AI systems are moving out of the lab and into the real world. For some applications, such as image recognition, AI systems have already beaten humans in benchmark tests. In other cases, such as language translation, current AI systems are not as effective as the best human translators but are good enough to be useful in some settings. AI systems may not need to achieve superhuman performance to be valuable, however. In some cases, their value may come from being cheaper, faster, or easier to deploy at scale relative to human expertise. Some examples of AI uses include:

- ***Data classification*** – AI systems can be used to help classify data, from images to song genres to medical imagery and diagnosis.[27] In many cases, AI systems can classify data more reliably and accurately than humans.

Real-time strategy games, such as Dota 2, are the latest frontier for AI systems in achieving superhuman performance in games because of their complexity and open-ended game play options. (OpenAI)

- ***Anomaly detection*** – AI systems can help detect anomalous behavior, such as fraudulent financial transactions or new malware.[28] AI systems can find anomalies whose signatures are not yet known by analyzing routine patterns of behavior (financial, cyber, or other) and then identifying new behavior that is outside the norm. These systems can be used to monitor large data streams at scale and in real time, in ways that would not be feasible for humans.
- ***Prediction*** – By finding patterns across large sets of data, AI systems can make statistical predictions about future behavior. Systems of this type are already in routine commercial use, such as search engine auto-fills and Netflix and Amazon recommendations. Machine learning has also demonstrated value in improving weather forecasting.[29] Some applications raise thorny ethical issues, such as using AI for predictive policing or estimating longevity of medical patients in end-of-life care.[30]
- ***Optimization*** – AI systems can be used to optimize performance for complex systems and tasks. For example, DeepMind used machine learning to optimize Google data centers to improve energy efficiency, resulting in a 40 percent savings in the amount of energy needed for cooling and a 15 percent overall improvement in energy efficiency.[31]

Autonomy

Artificial intelligence also allows the creation of machines with greater autonomy, or freedom, to perform tasks on their own. As machines become more capable, humans may be comfortable delegating them greater autonomy in a wider variety of settings. Autonomy has many advantages, including:

- ***Embedded expertise*** – Automation allows lesser-skilled individuals to perform tasks at or near the same level as higher-skilled workers by embedding expertise within the machine. For example, individuals can use tax preparation software to do their taxes, with the computer performing many of the tasks normally performed by an accountant. This can lower the barrier to entry for humans to perform certain tasks.
- ***Larger scale operations*** – Because software can be replicated at close to zero cost, automation allows the deployment of expertise at large scales. Tasks that normally could be done by humans, but only at small scales, can become feasible at larger scales with automation. Examples including automated spear phishing for cyber attacks, targeting advertising and sales to certain groups, and automated bug finding for discovering cyber vulnerabilities in software.
- ***Faster-than-human reaction times*** – Automation can perform tasks at superhuman speeds, reacting to events far more quickly than would be possible for humans. This is already the case today for high-frequency stock trading, which occurs in milliseconds, and automatic braking in automobiles.
- ***Superhuman precision and reliability*** – Automation can be used to perform many tasks with precision and reliability that surpasses human performance. The X-47B experimental drone demonstrated a degree of precision in its landings that would be impossible for humans to match. Robot-assisted surgery is used to perform miniaturized, high-precision surgery that is not possible with human hands.
- ***Superhuman patience and vigilance*** – Automated systems can monitor data without tiring or losing attention, keeping a vigilant eye on nuclear power plants or observing computer network activity for malware signatures.
- ***Operations without connections to humans*** – Autonomy enables robotic systems to perform missions independently without reliable communications to humans. In some cases this could be for long periods of time, such as autonomous underwater gliders that operate at sea for months at a time performing oceanographic surveys.[32]

Limitations of Current AI Systems

In spite of these advantages, artificial intelligence still has significant limitations. Current AI systems generally lack the ability to understand the context for their behavior, flexibly adapt to novel circumstances outside the parameters of their design, or employ what humans might think of as "common sense." A contemporary image recognition system, for example, could accurately identify objects in a scene, but would generally struggle to tell a coherent story about what was happening. Similarly, AI systems can accurately identify human faces and emotions and precisely track body movements, but would not be able to tell a plausible story explaining the motivations for a person's behavior. The result is an "idiot-savant" form of intelligence; AI systems may perform far better than humans in some areas while simultaneously failing to exhibit common sense.

AI Safety Concerns and Vulnerabilities

In addition to the general limitation of narrowness, current AI systems have a number of vulnerabilities and safety concerns that decision-makers should take into account. These are especially important for national security applications, where the consequences of mistakes or adversary hacking could be severe. Below is a brief overview of some of these problems.

Brittleness

The narrow nature of current AI systems can make their intelligence "brittle." Without the ability to understand the broader context for their actions, AI systems may not understand when that context has changed and their behavior is no longer appropriate. In constrained settings such as games, this can be less of an issue. In real-world settings, however, this means that AI systems can suddenly and dramatically fail if the environment or context for their use changes. They can move from super smart to super dumb in an instant. This can be true even for learning systems. Thus, human oversight and judgment in the deployment of AI systems is necessary to avoid or mitigate the risk of brittleness. Humans who supervise the operation of AI systems can step in to halt or change the operation of the system if the environment changes and it begins to fail.

AI systems may perform far better than humans in some areas while simultaneously failing to exhibit common sense.

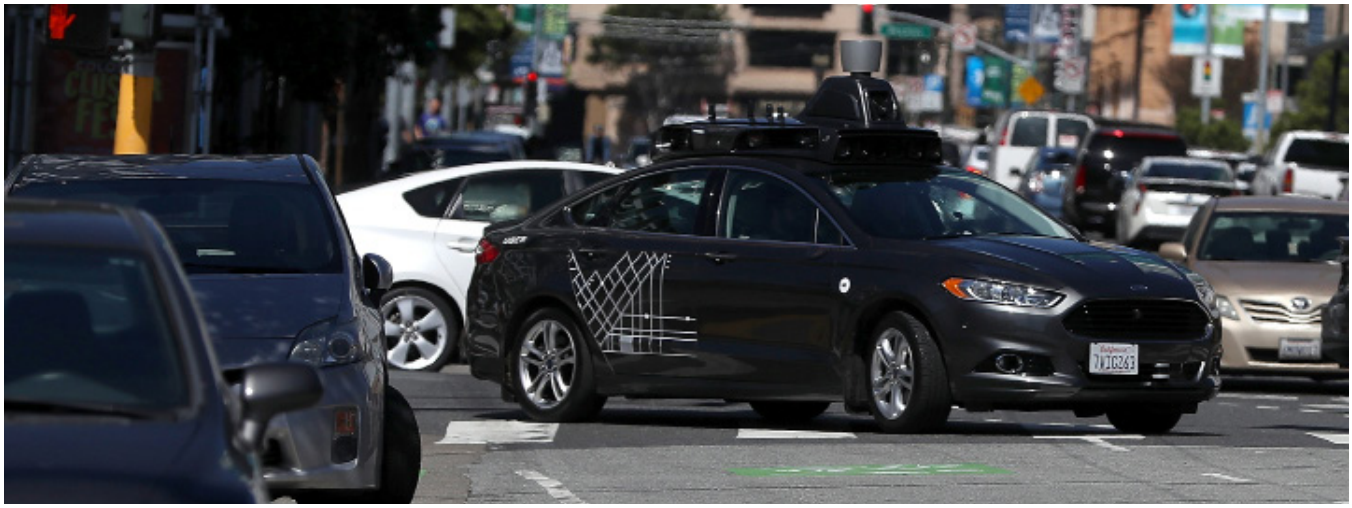
Predictability

Because of their complexity, it may not always be possible for users to anticipate the behavior of an AI system in advance. This means that users are sometimes surprised by how a system behaves. This can be exacerbated when AI systems are goal-oriented and/or interact with real-world environments. For example, a user may not be able to predict precisely when a self-driving car will change lanes or perform other maneuvers. Similarly, even the programmer who designed a chess-playing computer program may not be able to predict which moves the chess program will play. In both examples, given a set of general rules for how to behave, the AI system is given the authority to determine the best course of action to achieve a goal (driving to a destination; winning a chess game) based on the specific circumstances at the time (driving environment; location of pieces on the chess board). Rather than being a drawback, this flexibility is precisely the point of designing an AI system – to allow a machine to determine the best course of action to solve a problem, given a variety of potential environmental conditions. This feature of AI-enabled systems sometimes can be problematic, however, if the behavior of the system falls outside the bounds of the kinds of actions that the human user may expect or desire.

The problem of unpredictable behavior can occur even in systems that do not use machine learning. For example, in 2012 the financial trading firm Knight Capital Group was wiped out by a financial glitch that led their algorithms to execute 4 million erroneous trades in 45 minutes, resulting in a $460 million loss.[33] This happened even though Knight Capital Group's automated trading algorithms were relatively simple compared to today's cutting-edge AI methods. Better testing and evaluation of AI systems in realistic environments can help identify these behaviors in advance, but this challenge is likely to remain a risk for complex autonomous systems interacting with real-world environments.

Explainability

Some AI methods make it difficult, even after the fact, to explain the causes of their behavior. The behavior of rule-based systems is generally understandable, at least afterward, because a given behavior can be traced back to a particular rule or interaction of rules. For learning systems, the AI system's behavior depends on its prior experiences or training data. The information that deep neural networks use to identify images is encoded within the strength of connections within the network, for example, not a set of explainable rules. An AI image recognition system may be able to correctly identify an image of a school bus, but not be able to explain which features of the image cause it to conclude that the picture is a bus. This "black box" nature of AI systems may create challenges for some applications. For instance, it may not be enough for a medical diagnostics AI to arrive at a diagnosis; doctors are likely to also want to know which indicators the AI is using to do so. Research into more explainable AI methods is thus critical to expanding potential applications for AI systems. [34]



An Uber self-driving car navigates the streets of San Francisco in 2017. Driving is challenging for AI systems because it takes place in an unstructured environment. National security applications of AI often have an additional layer of complexity, since adversaries are trying to hack, spoof, or manipulate AI systems. (Justin Sullivan/Getty Images)

Machine Learning Safety Problems and Vulnerabilities

Machine learning techniques are powerful, but have a number of potential safety problems that can arise from failures at any stage of the learning process. Learning systems can generate flawed outcomes if their goal, or objective function, is improperly specified. This can happen in two possible ways. One way is if the goal does not take into account some important factors and, therefore, pursuit of the goal results in some negative side effect or harm. In their paper, "Concrete Problems in AI Safety," Dario Amodei, a research scientist at OpenAI, and other AI researchers give the thought experiment of a cleaning robot that inadvertently damages the environment by knocking over a vase because it was not programmed to avoid doing so.[35]

In a real-world example, independent researchers have claimed that YouTube's algorithm for recommending additional videos pushes viewers to extremist content. These researchers hypothesize that the algorithm is designed to maximize viewer time spent on YouTube, and that the algorithm learned that more inflammatory content kept viewers watching longer. If true, this would be an example of the pursuit of a goal (maximize ad revenue) having an unintended negative side effect (increasing exposure to extremist content).[36]

A second way in which an improperly specified goal can cause problems is if the algorithm engages in reward hacking. This is when the machine learns a behavior that technically meets its goal but is not what the designer intended. The system has therefore "hacked" its reward function. To the human observer, this often looks like the AI system is finding a loophole to meet the letter of its goal, but not the intent. In a comprehensive survey from 50 AI researchers on "The Surprising Creativity of Digital Evolution," the authors note:

> [I]t is often functionally simpler for evolution to exploit loopholes in the quantitative measure than it is to achieve the actual desired outcome. ... We often ascribe creativity to lawyers who find subtle legal loopholes, and digital evolution is often frustratingly adept at similar trickery.[37]

Examples of reward hacking abound from reinforcement learning and evolutionary algorithms in games and other digital simulation environments. Just a few examples include:

- A Tetris-playing bot learned to pause the game before the last brick fell so that it would never lose.[38]
- Simulated digital creatures evolved clever ways of falling to achieve their movement goals without actual locomotion or jumping.[39]
- In a naval strategy game that developed new rules for combat tactics, the top-scoring rule was one that learned to take credit for other rules.[40]
- A reinforcement learning system learned that the optimal scoring strategy in a boat racing game was not to race at all but to perform tight loops through auto-renewing targets mid-course, racking up more points than was possible from completing the race.[41]

- A computer program deleted the files containing the "correct" answers against which it was being evaluated, causing it to be awarded a perfect score.[42]

These safety problems raise questions about potential national security applications of machine learning. For example, a cybersecurity system tasked with defending networks against malware could learn that humans are a major source of introducing malware and lock them out (negative side effect). Or it could simply take a computer offline to prevent any future malware from being introduced (reward hacking). While these steps might technically achieve the system's goals, they would not be what the designers intended.

Safety problems can also arise from the data machines use to learn. AI systems can suffer from the same challenges of overfitting as statistical models in general. Data overfitting is when an AI system learns to mimic the training data precisely, rather than the underlying concepts the training data represents, so the system fails when applied outside of the training data. Another problem is ensuring robustness to changes in the input data. Even if a system is properly trained on an initial set of data, if the actual environment changes, the system may not be able to adequately adapt. This can be a common problem if the data used to train a learning system does not adequately represent the data that it will face in the real world. In one real-world example, parents posted a video of the voice-activated home computer device Alexa searching for adult content after hearing a toddler request a children's song.[43] A more robust set of training data that included small children, who may not pronounce words as clearly as adults, might have prevented this particular example, but this is a general problem for learning systems shifting from training datasets to interactions in the real world.

Even if a system is properly trained on an initial set of data, if the actual environment changes, the system may not be able to adequately adapt.

Bias

Bias – a deviation from a standard – can arise in AI systems in a variety of ways. Bias is not always problematic, but can be in some cases. One way an AI system can exhibit bias is if the objective function, or goal, mirrors a bias (explicit or implicit) on the part of the designers. If the AI system's objective accurately reflects the values of its designers, then in one sense it is a well-designed system. But if those goals are not socially desirable, then there could be harmful consequences to using the system. For example, a self-driving car that was programmed to always obey the speed limit, even if it might be safer in some settings to drive above the speed limit with the flow of traffic, would exhibit a bias toward compliance with the law over passenger safety.

Another way that an AI system can exhibit bias is if the training data is biased in some way. Some forms of bias could have moral connotations, if for example the data captures the biases of the people who collected, assembled, or chose the data. In other cases, the training data could be biased in a more technical sense, with the training data not being a representative sample of the actual operating environment. For example, a chess-playing program that was trained on human moves

A reinforcement learning system playing a boat racing game developed a strategy of circling auto-renewing targets to maximize its score, rather than attempting to win the race. (OpenAI)

might have a bias toward playing in ways that are cognitively easier for humans. This bias could be valuable if the intent of the chess program is to mimic a human player. If the intent is to play the best possible game of chess, however, then this bias could be harmful.

Bias could be a concern for national security applications where the training data deviates from the actual operating environment. For example, it is difficult for militaries to realistically simulate actual war. This introduces the potential for systems to be biased, and potentially in a way that militaries do not discover until combat. The fog and friction of real war mean that there are a number of situations in any battle that it would be difficult to train an AI to anticipate. Thus, in an actual battle, there could be significant risk of an error.

System Accidents

AI is also vulnerable to system failures, stemming from complex interaction among elements of a system. System accidents, which are possible in any sufficiently complex system, are exacerbated in competitive environments where actors are not incentivized to share their algorithms with one another. When AI systems are involved, the interaction among different algorithms can lead to bizarre behavior, sometimes at superhuman speeds. In 2011, two automated pricing bots on Amazon got caught in a price war and escalated the price of a biology textbook to $23,698,655.93 (plus $3.99 shipping).[44] More consequential examples of this phenomenon are stock trading flash crashes, which remain a persistent problem across multiple financial markets. Financial regulators have mitigated the consequences of these flash crashes by installing "circuit breakers" that take stocks offline if the price moves too quickly. These automated circuit breakers allow financial regulators to monitor individual stock prices – and react within seconds – at a scale that would not be feasible for humans. Flash crashes continue to occur, however, with over 1,200 circuit breakers tripped across multiple markets in one day in 2015.[45]

In national security settings, unintended interactions could occur by AI systems trying to gain a competitive advantage on one another and taking actions that could be destructive or counterproductive. In settings where machines interact at superhuman speeds, such as in cyberspace or electronic warfare, these interactions could lead to harmful consequences before humans users can adequately respond.

In national security settings, unintended interactions could occur by AI systems trying to gain a competitive advantage on one another and taking actions that could be destructive or counterproductive.

Human-Machine Interaction Failures

Even when AI systems work perfectly, accidents can still occur if the user does not fully understand the system's limitations or the feedback that it gives. When these accidents occur, observers frequently blame the human user, but the true cause is a breakdown between human and machine. Examples include the 2009 Air France 447 crash, which killed all passengers onboard,[46] and the 2016 crash of a Tesla on auto-pilot, which killed the driver.[47] In these cases, the human operators of highly complex automated systems failed to understand the information the system was giving them, leading to tragic consequences. This is a particular challenge for national security applications in which the user of the system might be a different individual than its designer and therefore may not fully understand the signals the system is sending. This could be the case in a wide range of national security settings, such as the military, border security, transportation security, law enforcement, and other applications where the system's designer is not likely to be either the person who decides to field the system or the end-user.

Exploiting Machine Learning Vulnerabilities

Malicious actors who deliberately seek to subvert AI systems can potentially manipulate these AI safety problems, creating a new category of risks. Financial traders have exploited the behavior of trading algorithms to artificially manipulate stock prices.[48] Similarly, adversaries could learn how AI systems behave and exploit their weaknesses. Malicious actors could also subvert learning systems by poisoning their data during the learning process, so that they learn incorrect behavior. This could be done by gaining access to the training data and manipulating it in subtle ways to create behavioral flaws once the system is trained. Alternatively, for learning systems that interact with the real world, adversaries could feed the system data that causes it to learn incorrect behaviors. The Microsoft chatbot Tay learned

to parrot racist and anti-Semitic language after less than 24 hours on Twitter.[49]

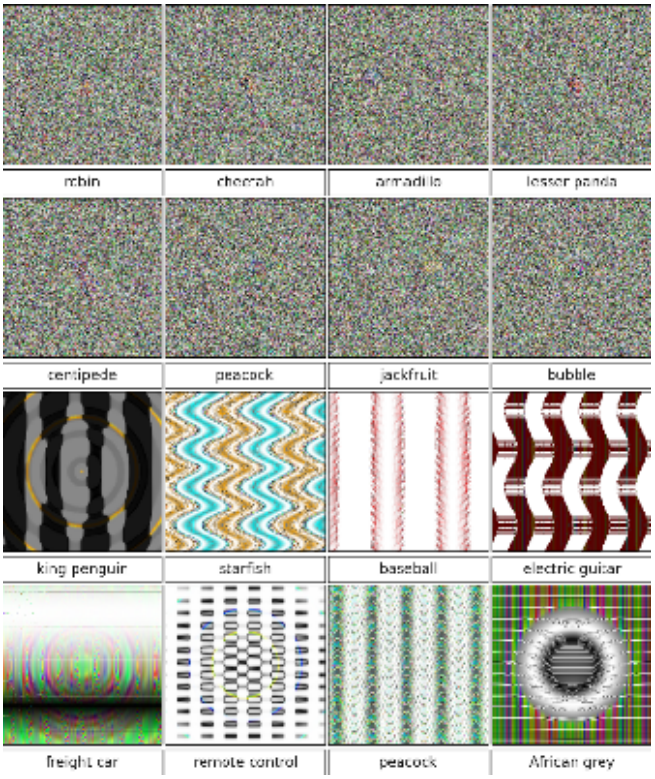
These vulnerabilities present significant challenges for artificial intelligence in national security applications, which often have high consequences for failure.[50] These problems are amplified in an inherently adversarial context in which both state and non-state actors will seek to exploit weaknesses in AI systems and manipulate their behavior.

These safety problems, which apply to a broad range of AI and machine learning techniques, are compounded by the vulnerability of deep neural networks to false data inputs (spoofing attacks). Neural networks that normally perform well at object classification tasks, such as image recognition, can be fooled by adversarial

Malicious actors who deliberately seek to subvert AI systems can potentially manipulate these AI safety problems, creating a new category of risks.

data. Adversarial inputs from a malicious actor, which to humans often look like random noise or nonsense images, can fool neural networks into believing the images are something else – and with high confidence.[51] These "fooling images" can even be embedded into other images in a way that is undetectable to humans. These attacks, which use specially generated adversarial data, can succeed even if the attackers do not have access to the training data or source code of the targeted neural network.[52] In such a case, a local model is trained based on the observed behavior of the neural network, such as classifying a particular image. Then, adversarial data that is inserted post-training to fool the local model is used to attack the original neural network. These fooling images can even be embedded into physical objects, in one demonstration causing an image classifier to misidentify a 3D-printed turtle as a rifle.

Despite significant research on the adversarial data problem, AI researchers do not yet have a workable solution to protect against this form of attack. Because of this vulnerability, image recognition systems could be fooled by counter-AI camouflage, causing the image recognition system to misidentify objects. Adversaries could make a tank look like a school bus, and vice versa. Even worse, these patterns could be hidden in a way that is undetectable by humans. Decoy objects could be scattered around the environment, confusing neural



Specially evolved "fooling images" can be fed into AI-based image classifiers to trick them into misidentifying images with high confidence. A neural network–based image classifier identified all of the above images as the associated labels with greater than 99 percent confidence. (Anh Nguyen, Jason Yosinski, and Jeff Clune)[54]

network–based sensors, and valid targets could be covered with camouflage designed to make them appear innocuous. AI researchers have demonstrated the ability to do this relatively easily – for example, making a stop sign appear to an image classifier to be a 45 mile per hour sign simply by adding some small black and white stickers. This form of passive environmental hacking could be done well in advance of an AI system scanning an environment, like a cognitive land mine waiting to fool a system when it arrives. Neural network–based data classifiers are likely too valuable to ignore, so national security decision-makers will need to factor in these vulnerabilities when using AI, whether for image recognition or other activities.

The Capability-Vulnerability Tradeoff

It may be tempting to assume that responsible actors will not employ AI and machine learning systems until these vulnerabilities are solved, but that is not likely to be the case. Computers are vulnerable to hacking, yet that has not stopped their use across society and in national security settings, even when data breaches have had

serious consequences.[53] The advantages of using computer network technology are too great to ignore, and AI systems are similarly attractive. AI systems are powerful and have many benefits, but are vulnerable to hacking that exploits weaknesses in how they learn, process data, and make decisions. These risks are heightened in national security settings, which are often adversarial, high consequence, and difficult to replicate in training environments. Policymakers must be aware of these risks and seek to mitigate against these vulnerabilities as much as possible in the design and use of AI systems.

Future AI Progress

The field of AI and machine learning has advanced dramatically in only the past few years and continues to move forward in leaps and bounds. The future of AI is highly uncertain. One key variable is progress toward creating more general-purpose AI systems that could exhibit intelligent behaviors across multiple domains, unlike today's narrow AI systems. Another significant variable is progress on unsolved safety problems and vulnerabilities in AI systems. A world where AI performance outpaces safety could be quite hazardous if nations race to put into the field AI systems that are subject to accidents or subversion (e.g., spoofing attacks). On the other hand, progress in AI safety could mitigate some of the risks that stem from national security uses of AI. Much of the innovation in the field of artificial intelligence is being driven by the commercial sector, but governments do have the ability to influence the direction of progress through research investments. The U.S. government should increase its investment in AI safety to improve the prospects for building robust, reliable, and explainable AI systems in national security settings.

Endnotes

1. Kevin Kelly, "The Three Breakthroughs That Have Finally Unleashed AI on the World," Wired, October 27, 2014, https://www.wired.com/2014/10/future-of-artificial-intelligence/.

2. "Putin: Leader in artificial intelligence will rule the world," Associated Press, September 1, 2017, https://nypost.com/2017/09/01/putin-leader-in-artificial-intelligence-will-rule-the-world/.

3. Adapted from Nils J. Nilsson, The Quest for Artificial Intelligence: A History of Ideas and Achievements (Cambridge, U.K.: Cambridge University Press, 2010).

4. Adapted from Shane Legg and Marcus Hutter, "A Collection of Definitions of Intelligence," Preprint, submitted June 25, 2007, https://arxiv.org/pdf/0706.3639.pdf.

5. Adapted from Tom Michael Mitchell, "The Discipline of Machine Learning," Carnegie Mellon University, School of Computer Science, Machine Learning Department (2006); and Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers: What It Is and Why It Matters," (Belfer Center, June 2017), https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

6. "Artificial Intelligence Index: 2017 Annual Report," (AI Index, November 2017), 26, http://cdn.aiindex.org/2017-report.pdf; and Peter Eckersley et al., Electronic Frontier Foundation Artificial Intelligence Measurement Project, https://www.eff.org/ai/metrics#Vision.

7. Adapted from Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction, Vol. 1 (Cambridge: MIT Press, 1998).

8. Alex Krizhevksky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" (paper presented at the Neural Information Processing Systems Proceedings, 2012), https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

9. Ian J. Goodfellow et al., "Generative Adversarial Nets," Preprint, submitted June 10, 2014, https://arxiv.org/pdf/1406.2661.pdf.

10. Robert Chesney and Danielle Citron, "Deep Fakes: A Looming Crisis for National Security Democracy, and Privacy?," Lawfare blog, February 21, 2018, https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy; Samantha Cole, "We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now," Motherboard.Vice.com, January 24, 2018, https://motherboard.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley; Kevin Rose, "Here Come the Fake Videos, Too," The New York Times, March 04, 2018, https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html; and Terro Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation" (paper presented at the Sixth International Conference on Learning Representations, Vancouver, Canada, April 30, 2018-May 03, 2018), http://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf.

11. Stanford Vision Lab, Stanford University, Princeton University, "About ImageNet: Summary and Statistics," last modified 2016, http://image-net.org/about-stats; and Stanford Vision Lab, Stanford University, Princeton University, "About ImageNet: Overview," last modified 2016, http://image-net.org/about-overview.

12. David Silver et al., "Mastering the game of Go without Human Knowledge," (DeepMind, October 18, 2017), 355, https://deepmind.com/documents/119/agz_unformatted_nature.pdf.

13. Melvin Johnson et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," Preprint, submitted August 21, 2017, https://arxiv.org/pdf/1611.04558.pdf.; and Mike Schuster, "Zero-Shot Translation With Google's Multi-Lingual Neural Machine Translation System," Google Research blog, November 22, 2016, https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

14. David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," Preprint, submitted December 5, 2017, https://arxiv.org/pdf/1712.01815.pdf.

15. Lasse Espeholt et al., "IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures," Preprint, submitted February 9, 2018, https://arxiv.org/pdf/1802.01561.pdf.; and Hubert Soyer, Drew Purves, and Lasse Espeholt, "IMPALA: Scaleable Distributed DeepRL in DMLab-30," DeepMind blog, February 5, 2018, https://deepmind.com/blog/impala-scalable-distributed-deeprl-dmlab-30/.

16. Peter Stone et al. "Artificial Intelligence and Life in 2030," One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel (Stanford University, September 06, 2016), 13, https://ai100.stanford.edu/sites/default/files/ai100report10032016fnl_singles.pdf.

17. Jonathan Schaeffer et al., "Checkers Is Solved," Science, 317 no. 5844 (September 2007), 1518-1522.

18. John Tromp, "Number of Legal Go Positions," http://tromp.github.io/go/legal.html; and DeepMind, "AlphaGo," https://deepmind.com/research/alphago/.

19. Cade Metz, "Inside Libratus, the Poker AI That Outbluffed the Best Humans," Wired, February 1, 2017, https://www.wired.com/2017/02/libratus/; Will Knight, "Why Poker Is a Big Deal for Artificial Intelligence," MIT

Technology Review, January 23, 2017, https://www.technologyreview.com/s/603385/why-poker-is-a-big-deal-for-artificial-intelligence/; and Matej Moravčík et al. "DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker," Preprint, submitted March 03, 2017, https://arxiv.org/pdf/1701.01724.pdf.

20. Demis Hassabis, "AlphaGo: Using machine learning to master the ancient game of Go," Google blog, January 27, 2016, https://blog.google/topics/machine-learning/alphago-machine-learning-game-go/.

21. Demis Hassabis and David Silver, "AlphaGo Zero: Learning from scratch," DeepMind blog, October 18, 2017, https://deepmind.com/blog/alphago-zero-learning-scratch/.

22. David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," December 5, 2017, https://arxiv.org/pdf/1712.01815.pdf.

23. OpenAI, "Dota 2," OpenAI blog, August 11, 2017, https://blog.openai.com/dota-2/.

24. Oriol Vinyals, Stephen Gaffney, and Timo Ewalds, "DeepMind and Blizzard open StarCraft II as an AI research environment," DeepMind blog, August 9, 2017, https://deepmind.com/blog/deepmind-and-blizzard-open-starcraft-ii-ai-research-environment/; and Yoochul Kim and Minhyung Lee, "Humans Are Still Better Than AI at StarCraft–For Now," MIT Technology Review, November 1, 2017, https://www.technologyreview.com/s/609242/humans-are-still-better-than-ai-at-starcraftfor-now/.

25. Kai Arulkumaran et al., "A Brief Survey of Deep Reinforcement Learning," Preprint, submitted September 28, 2017, https://arxiv.org/pdf/1708.05866.pdf.

26. OpenAI, "More on Dota 2," OpenAI blog, August 16, 2017, https://blog.openai.com/more-on-dota-2/.

27. Christopher Mims, "Using Neural Networks to Classify Music," MIT Technology Review, June 3, 2010, https://www.technologyreview.com/s/419223/using-neural-networks-to-classify-music/; Tom LH. Li, Antoni B. Chan, and Andy HW. Chun, "Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network" (paper presented at the proceedings of the International MultiConference of Engineers and Computer Scientists 2010, Hong Kong, March 17-19, 2010), http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp546-550.pdf; and Dave Fornell, "How Artificial Intelligence Will Change Medical Imaging," ImagingTechnologyNews.com, February 24, 2017, https://www.itnonline.com/article/how-artificial-intelligence-will-change-medical-imaging.

28. Efstathios Kirkos, Charalambos Spathis, and Yannis Monolopoulous, "Data Mining Techniques for the detection of fraudulent financial statements," Expert Systems with Applications, 32 (2007), 995-1003, http://delab.csd.auth.gr/papers/ESWA07ksm.pdf; and "DeepArmor: A cognitive approach to endpoint protection," SparkCognition.com, https://www.sparkcognition.com/deeparmor-enterprise/.

29. Emilcy Hernández et al., "Rainfall Prediction: A Deep Learning Approach" (paper presented at the proceedings of the International Conference on Hybrid Artificial Intelligence Systems, Seville, Spain, April 18-20, 2016), https://link.springer.com/chapter/10.1007/978-3-319-32034-2_13.

30. Randy Rieland, "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?," Smithsonian.com, March 5, 2018, https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/; and Anand Avanti et al., "Improving Palliative Care with Deep Learning," Preprint, submitted November 17, 2017, https://arxiv.org/pdf/1711.06402.pdf.

31. Richard Evans and Jim Gao, "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%," DeepMind blog, July 20, 2016, https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/.

32. "Spray Glider," Woods Hole Oceanographic Institution, https://www.whoi.edu/main/spray-glider.

33. D7, "Knightmare: A DevOps Cautionary Tale," Doug Seven, April 17, 2014, https://dougseven.com/2014/04/17/knightmare-a-devops-cautionary-tale/.

34. David Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/explainable-artificial-intelligence; David Gunning, "Explainable Artificial Intelligence (XAI)," Lecture, https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf; Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, "Explaining Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," Preprint, submitted August 28, 2017, https://arxiv.org/pdf/1708.08296.pdf; Andreas Holzinger et al., "What do we need to build explainable AI systems for the medical domain?," Preprint, submitted December 28, 2017, https://arxiv.org/pdf/1712.09923.pdf; and Lisa Anne Hendricks et al., "Generating Visual Explanations," Preprint, submitted March 28, 2016, https://arxiv.org/pdf/1603.08507.pdf.

35. Amodei et al., "Concrete Problems in AI Safety," Preprint, submitted July 25, 2016, 4, https://arxiv.org/pdf/1606.06565.pdf.

36. This problem is not unique to machine learning and can arise in any goal-oriented system that has an improperly specified goal.

37. Joel Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," Preprint, submitted March 8, 2018, 6, https://arxiv.org/pdf/1803.03453.pdf.

38. Tom Murphy VII, "The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel … after that it gets a little tricky," https://www.cs.cmu.edu/~tom7/mario/mario.pdf; Douglas B. Lenat, "EURISKO: A Program That Learns New Heuristics and Domain Concepts," Artificial Intelligence 21 (1983), http://www.cs.northwestern.edu/~mek802/papers/not-mine/Lenat_EURISKO.pdf, 90; and Jack Clark and Dario Amodei, "Faculty Reward Functions in the Wild," OpenAI blog, December 21, 2016, https://blog.openai.com/faulty-reward-functions/.

39. Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," 6-7.

40. Lenat, "EURISKO: A Program That Learns New Heuristics and Domain Concepts."

41. Dario Amodei and Jack Clark, "Faulty Reward Functions in the Wild," OpenAI blog, December 21, 2016, https://blog.openai.com/faulty-reward-functions/.

42. Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," 7.

43. Entrepreneur Staff, "Whoops, Alexa Plays Porn Instead of a Kids Song!," Entrepreneur.com, January 3, 2017, https://www.entrepreneur.com/video/287281.

44. Michael Eisen, "Amazon's $23,698,655.93 book about flies," It is Not Junk, April 22, 2011, http://www.michaeleisen.org/blog/?p=358.

45. Matt Egan, "Trading Was Halted 1,200 Times Monday," CNNMoney, August 24, 2015, http://money.cnn.com/2015/08/24/investing/stocks-markets-selloff-circuit-breakers-1200-times/index.html; and Todd C. Frankel, "Mini flash crash? Trading anomalies on Manic Monday hit small investors," The Washington Post, August 26, 2015, https://www.washingtonpost.com/business/economy/mini-flash-crash-trading-anomalies-on-manic-monday-hit-small-investors/2015/08/26/6bdc57b0-4c22-11e5-bfb9-9736d04fc8e4_story.html?utm_term=.749eb0bbbf5b.

46. "Final Report: On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janiero – Paris," (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, July 2012), [English translation], http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf; William Langewiesche, "The Human Factor," Vanity Fair, October 2014, http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash; and Nick Ross and Neil Tweedie, "Air France Flight 447: 'Damn it, we're going to crash,'" The Telegraph, April 28, 2012, http://www.telegraph.co.uk/technology/9231855/Air-France-Flight-447-Damn-it-were-going-to-crash.html.

47. Jim Puzzanghera, "Driver in Tesla crash relied excessively on Autopilot, but Tesla shares some blame, federal panel finds," The Los Angeles Times, September 12, 2017, http://www.latimes.com/business/la-fi-hy-tesla-autopilot-20170912-story.html; "Driver Errors, Overreliance on Automation, Lack of Safeguards, Led to Fatal Tesla Crash," National Transportation Safety Board Office of Public Affairs, press release, September 12, 2017, https://www.ntsb.gov/news/press-releases/Pages/PR20170912.aspx; and "Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida" NTSB/HAR-17/02/PB2017-102600 (National Transportation Safety Board, May 7, 2016), https://www.ntsb.gov/news/events/Documents/2017-HWY16FH018-BMG-abstract.pdf.

48. "Futures Trader Pleads Guilty to Illegally Manipulating the Futures Market in Connection With 2010 'Flash Crash,'" Department of Justice Office of Public Affairs, press release, November 9, 2016, https://www.justice.gov/opa/pr/futures-trader-pleads-guilty-illegally-manipulating-futures-market-connection-2010-flash.

49. James Vincent, "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day," TheVerge.com, March 24, 2016, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist; and Sophie Kleeman, "Here Are the Microsoft Twitter Bot's Craziest Racist Rants," Gizmodo.com, March 23, 2016, https://gizmodo.com/here-are-the-microsoft-twitter-bot-s-craziest-racist-ra-1766820160.

50. There are other important safety problems in machine learning systems, including safe exploration and scalable oversight. For a comprehensive evaluation of AI safety problems, see: Amodei et al., "Concrete Problems in AI Safety," Preprint, submitted July 25, 2016, https://arxiv.org/pdf/1606.06565.pdf.

51. Anh Nguyen A et al., "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

52. Nicolas Papernot et al., "Practical Black-Box Attacks against Machine Learning," Preprint, submitted March 19, 2017, https://arxiv.org/pdf/1602.02697.pdf.

53. Wikipedia contributors, "List of data breaches," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=List_of_data_breaches&oldid=831831575.

54. Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts and the public with innovative, fact-based research, ideas and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, and was established in February 2007 by co-founders Kurt M. Campbell and Michèle A. Flournoy.

CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and non-partisan. CNAS does not take institutional positions on policy issues. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the authors.

1152 15th Street, NW Suite 950 Washington, DC 20005
t. 202.457.9400 | f. 202.457.9401 | info@cnas.org | cnas.org

Bold. Innovative. Bipartisan.