Understanding AI Technology

A concise, practical, and readable overview of Artificial Intelligence and Machine Learning technology designed for non-technical managers, officers, and executives
April 2020

By: Greg Allen, Chief of Strategy and Communications
Joint Artificial Intelligence Center (JAIC)
Department of Defense

Foreword by JAIC Director Lt Gen Jack Shanahan

Acknowledgments
The author would like to thank the following individuals for their assistance reviewing earlier drafts of this document:

- Dr. Jeff Alstott (IARPA)
- Dr. Nate Bastian (Major, U.S. Army, DoD Joint AI Center)
- Dr. Steven L. Brunton (University of Washington)
- Dr. Matthew Daniels (Georgetown University)
- Dr. Ed Felten (Princeton University)
- Mr. Rob Jasper (Pacific Northwest National Laboratory)
- Dr. John Launchbury (Galois, and formerly DARPA)

Disclaimer
The views expressed in this document are those of the author alone and do not necessarily reflect the position of the Department of Defense or the United States Government.

Website: https://www.ai.mil/
Twitter: @DoDJAIC
LinkedIn: https://www.linkedin.com/company/dod-joint-artificial-intelligence-center/

FOREWORD BY LIEUTENANT GENERAL JACK SHANAHAN

It is hard for me to describe the steep slope of the learning curve I faced when I started the Project Maven journey over three years ago. While in many ways I still consider myself an Artificial Intelligence neophyte today, what I knew about the subject back then could barely fill the first few lines of a single page in my trusty notebook. My journey of discovery since then has been challenging, to say the least. I only wish Greg Allen's guide to "Understanding AI Technology" had been available to me in late 2016 as we embarked on our first AI/ML pilot project for ISR full-motion video analysis.

Greg has performed an inestimable service by writing this guide. AI is changing national security, and it's essential that DoD leaders have a firm grasp of the technology's building blocks. As I learned back in 2017 and am reminded daily in my role as the Director of the Joint AI Center (JAIC), AI is not an elixir. It is an enabler – one that is critical to our future national security. It is important for all of us to share the same fundamental understanding of AI technology. Greg's guide balances breadth and depth in just the right way. It is clear, concise, and cogent. I am confident it will be a valuable resource for everyone in DoD and beyond.

Lieutenant General John N.T. "Jack" Shanahan

Director, Joint Artificial Intelligence Center
Department of Defense
April 2020

EXECUTIVE SUMMARY

Many officials throughout the Department of Defense are asked to make decisions about AI before they have an appropriate understanding of the technology's basics. This guide will help.

The DoD AI Strategy defines AI as "the ability of machines to perform tasks that normally require human intelligence." This definition includes decades-old DoD AI, such as aircraft autopilots, missile guidance, and signal processing systems. Though many AI technologies are old, there have been legitimate technological breakthroughs over the past ten years that have greatly increased the diversity of applications where AI is practical, powerful, and useful. Most of the breakthroughs and excitement about AI in the past decade have focused on Machine Learning (ML), which is a subfield of AI. Machine Learning is closely related to statistics and allows machines to learn from data.

The best way to understand Machine Learning AI is to contrast it with an older approach to AI, Handcrafted Knowledge Systems. Handcrafted Knowledge Systems are AI that use traditional, rules-based software to codify subject matter knowledge of human experts into a long series of programmed "if given x input, then provide y output" rules. For example, the AI chess system Deep Blue, which defeated the world chess champion in 1997, was developed in collaboration between computer programmers and human chess grandmasters. The programmers wrote (literally typed by hand) a computer code algorithm that considered many potential moves and countermoves and reflected rules for strong chess play given by human experts.

Machine Learning systems are different in that their "knowledge" is not programmed by humans. Rather, their knowledge is learned from data: a Machine Learning algorithm runs on a training dataset and produces an AI model. To a large extent, Machine Learning systems program themselves. Even so, humans are still critical in guiding this learning process. Humans choose algorithms, format data, set learning parameters, and troubleshoot problems.

Machine Learning has been around a long time, but it previously was almost always expensive and complicated with low performance, so there were comparatively few applications and organizations for which it was a good fit.

Thanks to the ever-increasing availability of massive datasets, massive computing power (both from using GPU chips as accelerators and from the cloud), open source code libraries, and software development frameworks, the performance and practicality of using Machine Learning AI systems has increased dramatically.

There are four different families of Machine Learning algorithms, which differ based on aspects of the data they train on. It is important to understand the different families because knowing which family an AI system will use has implications for effectively enabling and managing the system's development.

1) Supervised Learning uses example data that has been labeled by human "supervisors." Supervised Learning has incredible performance, but getting sufficient labeled data can be difficult, time-consuming, and expensive.
2) Unsupervised Learning uses data but doesn't require labels for the data. It has lower performance than Supervised Learning for many applications, but it can also be used to tackle problems where Supervised Learning isn't viable.
3) Semi-Supervised Learning uses both labeled and unlabeled data and has a mix of the pros and cons of Supervised and Unsupervised learning.
4) Reinforcement Learning has autonomous AI agents that gather their own data and improve based on their trial and error interaction with the environment. It shows a lot of promise in basic research, but so far Reinforcement Learning has been harder to use in the real world. Regardless, technology firms have many noteworthy, real-world success stories.

Deep Learning (Deep Neural Networks) is a powerful Machine Learning technique that can be applied to any of the four above families. It provides the best performance for many applications. However, the technical details are less important for those not on the engineering staff or directly overseeing the procurement of these systems. What matters most for program management is whether or not the system uses Machine Learning, and whether or not the selected algorithm requires labeled data.

Systems using Machine Learning software can provide very high levels of performance. However, Machine Learning software has failure modes – both from accidents and from adversaries – that are distinct from those of traditional software. Program managers, system developers, test and evaluation personnel, and system operators all need to be familiar with these failure modes to ensure safe, secure, and reliable performance of AI systems.

There are multiple steps to developing an operational Machine Learning AI system. Usually, the biggest challenges relate to getting sufficient high-quality training data. System performance is directly tied to data quantity, quality, and representativeness.

Organizations should not pursue using AI for its own sake. Rather, they should have specific metrics for organizational performance and productivity that they are seeking to improve. Merely developing a high-performing AI model will not by itself improve organizational productivity. The model has to be integrated into operational technology systems, organizational processes, and staff workflows. Almost always, there will be some changes needed to existing processes to take full advantage of the AI model's capability. Adding AI technology without revising processes will deliver only a tiny fraction of the potential improvements, if any. Finally, traditional project management wisdom still applies. Many AI projects fail not because of the technology, but because of a failure to properly set expectations, integrate with legacy systems, and train operational personnel.

Purpose: By now, nearly all DoD officials understand that the rise of AI is an important technology trend with significant implications for national security, but many struggle to give simple and accurate answers to basic questions like:

- What is AI?
- How does AI work?
- Why is now an important time for AI?
- What are the different types of Machine Learning? How do they differ?
- What are Neural Networks and Deep Learning?
- What are the steps of building and operating AI systems?
- What are the limitations and risks of using of AI systems?

Contrary to popular belief, you do not need to understand advanced mathematics or know computer programming languages to be able to answer the above questions accurately and to develop a practical understanding of AI relevant to your organization's needs. This guide will cover everything that the vast majority of DoD leaders need to know.

WHAT IS AI?

The DoD AI Strategy states that "AI refers to the ability of machines to perform tasks that normally require human intelligence." This definition is so obvious that many are confused by its simplicity. In fact, however, this definition is very similar to the ones used by many leading AI textbooks and leading researchers. The first thing to note about this definition is that AI is an extremely broad field, one that covers not only the breakthroughs of the past few years, but also the achievements of the first electronic computers dating back to the 1940s.

The definition of "Artificial Intelligence," is a bit of a moving target. When something is new and exciting, people have no qualms about labeling it "Artificial Intelligence." Once the capabilities of a particular AI approach are familiar, though, they are often called merely "software." This paper will return later to the subject of how modern AI approaches are different and why now is a critical moment for AI technology. For now, just understand that even old technology can still be AI.

HOW DOES AI WORK?

AI is defined by a set of capabilities, rather than a specific technical approach to achieving those capabilities. There are many different approaches to developing AI systems, and various approaches work differently with different strengths and weaknesses. DARPA, a longtime pioneer in AI research, has helpfully grouped many of these approaches into two broad categories: (1) Handcrafted Knowledge and (2) Machine Learning. Machine Learning systems are the newer of the two approaches (though still decades old) and are responsible for the dramatic improvements in AI capabilities over the past ten years. If you've heard

some person or company claim that their system "uses AI," most likely they mean that their system is using Machine Learning, which is a far cry from their system being an autonomous intelligence equal to or greater than human intellect in all categories. Still, recent progress in Machine Learning is a big deal, with implications for nearly every industry, including defense and intelligence. The easiest way to understand Machine Learning systems, however, is by contrasting them with Handcrafted Knowledge systems, so this paper will begin there.

Handcrafted Knowledge AI

Handcrafted Knowledge Systems are the older of the two AI approaches, nearly as old as electronic computers. At their core, they are merely software developed in cooperation between computer programmers and human domain subject matter experts. Handcrafted Knowledge Systems attempt to represent human knowledge into programmed sets of rules that computers can use to process information. In other words, the "intelligence" of the Handcrafted Knowledge System is merely a very long list of rules in the form of "if given x input, then provide y output." When hundreds or thousands or millions of these domain-specific rules are combined successfully – into "the program" – the result is a machine that can seem quite smart and can also be very useful.

A well-known example of a Handcrafted Knowledge AI System in widespread use is tax preparation software. By requiring users to input their tax information according to pre-specified data formats and then processing that data according to the formally programmed rules of the tax code (developed in cooperation between human software engineers and accountants), the output can be good enough to pass an IRS audit. When first introduced in the 1980s, tax preparation software was very successfully marketed as Artificial Intelligence. Now that it has been in widespread use for decades, however, calling it "Artificial Intelligence" has fallen out of fashion. Nevertheless, it still falls within both the DoD definition of AI and the formal definition used by most researchers in the field.

Another famous example of a Handcrafted Knowledge System is "Deep Blue," the IBM-developed, chess-playing AI that defeated the human world chess champion in 1997. Deep Blue was developed in cooperation between IBM's software engineers and several chess grandmasters, who helped translate their human chess expertise into tens of thousands of computer code rules for playing grandmaster-level chess.

Both tax preparation AI systems and Deep Blue are a specific type of Handcrafted Knowledge AI known as an Expert System. Another type of Handcrafted Knowledge AI is a Feedback Control System, which uses human-authored rules to compute system output based on sensor measurement inputs. Feedback Control Systems have been in widespread use by the Department of Defense for decades. Aircraft autopilots, missile guidance systems, and electromagnetic

signal processing systems are just a few of the thousands of high-performing, extremely reliable Feedback Control AI Systems that the Department of Defense and its partners have developed and operated over the past eight decades. In this sense, Handcrafted Knowledge Systems are victims of their own success. They are so common that they are generally no longer referred to as "AI" in common discourse. Nevertheless, Handcrafted Knowledge Systems remain important and useful. In some areas, such as tax preparation, they still have far higher performance than Machine Learning systems. In other areas, such as chess playing, language translation, and image classification, Handcrafted Knowledge AI systems have been greatly surpassed in performance by Machine Learning AI systems. Regardless, Handcrafted Knowledge systems will continue to improve and see wide use for decades to come.

Machine Learning AI

The key difference between a Handcrafted Knowledge System and a Machine Learning system is in where it receives its knowledge. Rather than having their knowledge be provided by humans in the form of hand-programmed rules, Machine Learning systems generate their own rules. For Machine Learning systems, humans provide the system training data. By running a human-generated algorithm on the training dataset, the Machine Learning system generates the rules such that it can receive input *x* and provide correct output *y*.

In other words, the system learns from examples (training data), rather than being explicitly programmed. This is why data is so vital in the context of AI. Data is the main raw material out of which high-performing Machine Learning AI systems are built. For this reason, the quality, quantity, representativeness and diversity of data will directly impact the operational performance of the ML system. Algorithms and computing hardware are also important, but nearly all ML systems run on commodity computing hardware, and nearly all of the best algorithms are freely available worldwide. Hence, having enough of the right data tends to be the key.

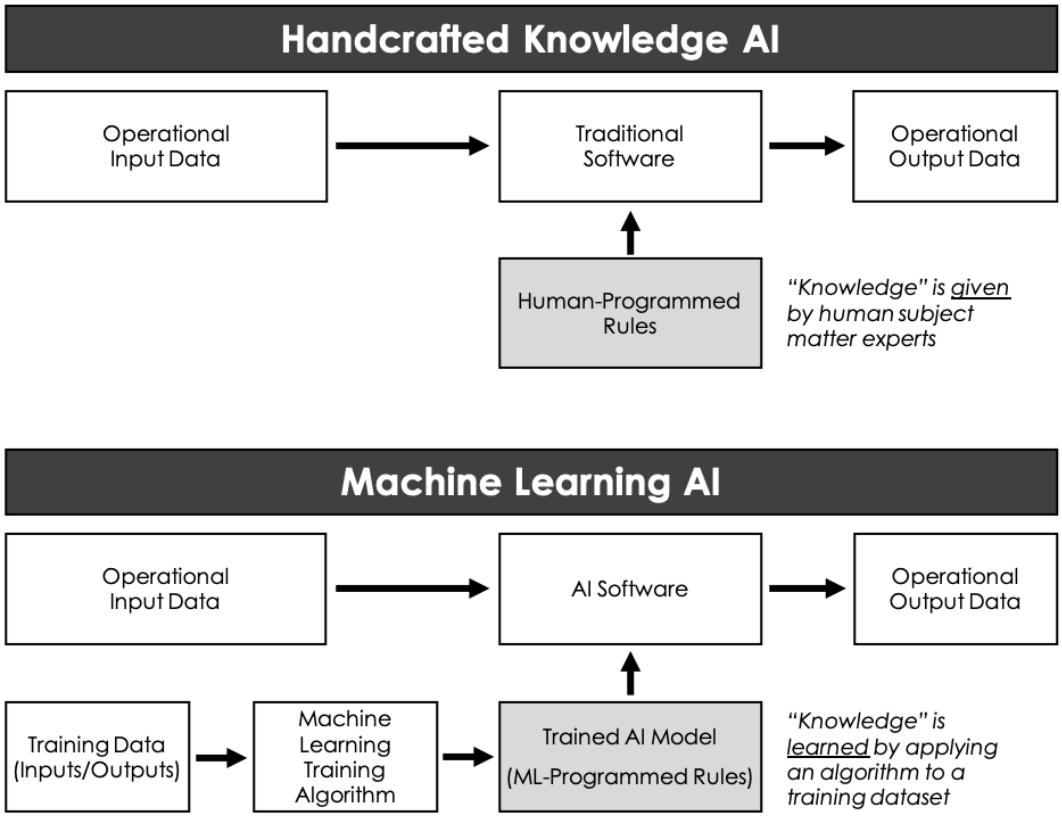
While it is true that – to a large extent – Machine Learning systems program themselves, humans are still critical in guiding this learning process: humans choose algorithms and datasets, format data, set learning parameters, and troubleshoot problems.

At this stage, many readers may ask themselves, "so what? Why is Machine Learning important?"

The reason is that there are many applications where task automation would be useful, but where human programming of all of the software rules to implement automation is either impractical or genuinely impossible. Sometimes human experts are unable to fully translate their intuition decision-making into fixed rules. Further, for a surprisingly large subset of these applications, the performance of

Machine Learning systems is very high, much higher than was ever achieved with Handcrafted Knowledge Systems or indeed by human experts. This does not mean that Handcrafted Knowledge systems are obsolete. For many applications they remain the cheapest and/or highest performing approach.

Figure 1: Simplified Diagram of AI Approaches



Pattern recognition, image analysis, language translation, content generation, and speech transcription are just a few noteworthy examples where the past performance of Handcrafted Knowledge AI was very low, but the performance of Machine Learning AI is extremely high, often better than human performance. Because of the increased performance and enhanced productivity Machine Learning enables, there are many practical applications throughout the economy and industry. We are not after using AI for its own sake. We are after increased performance and enhanced productivity. It's that simple.

WHY IS NOW AN IMPORTANT TIME FOR AI?

AI has been around for decades. So, why has everyone been talking about it constantly in recent years? It boils down to this – for Machine Learning AI systems – there has been a massive increase in the number of real-world applications where AI is now practical and powerful. There are four main reasons why this is true now but was not true ten years ago:

1) **More Massive Datasets:** Machine Learning algorithms tend to require large quantities of training data in order to produce high performance AI models. For example, some facial recognition AI systems can now routinely outperform humans, but to do so requires tens of thousands or millions of labeled images of faces for training data. When Machine Learning was first developed decades ago, there were very few applications where sufficiently large training data was available to build high performance systems. Today, an enormous number of computers and digital devices and sensors are connected to the internet, where they are constantly producing and storing large volumes of data, whether in the form of text, numbers, images, audio, or other sensor data files.

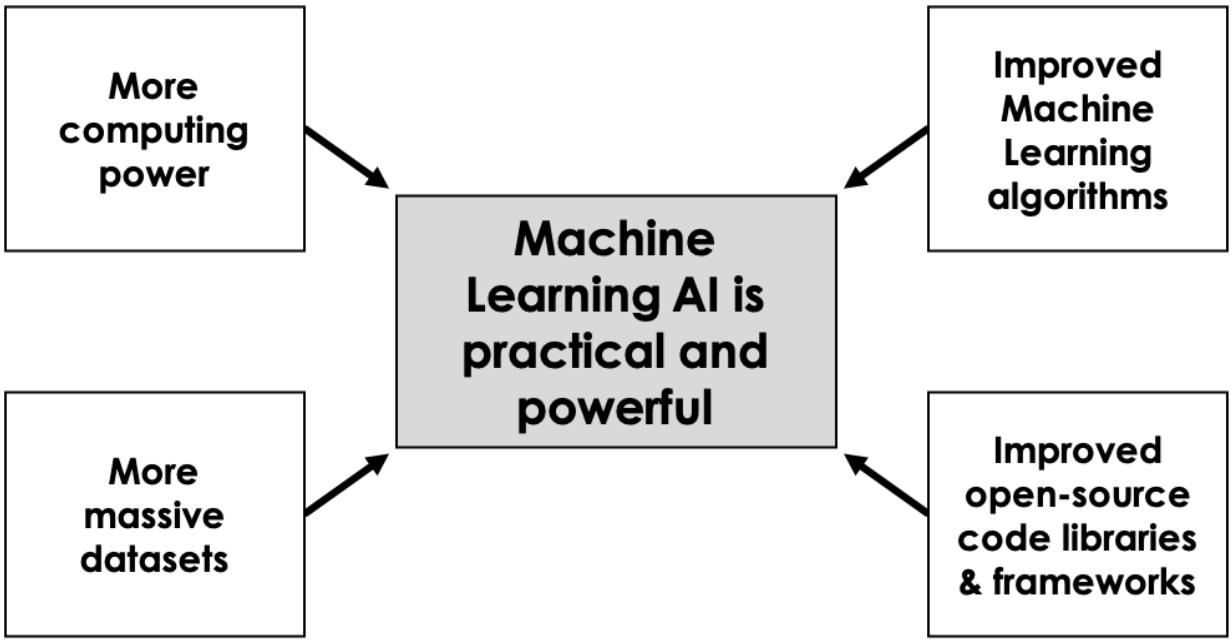
   Of course, more data only helps if the data is relevant to your desired application. If you're trying to develop a better aircraft autopilot, then a bunch of consumer loan application data isn't going to help, no matter how much you have. In general, training data needs to match the real-world operational data very, very closely to train a high-performing AI model.

2) **Increased Computing Power**: To a far greater extent than Handcrafted Knowledge Systems, Machine Learning AI systems require a lot of computing power to process and store all the above-mentioned data. Around ten years ago, computing hardware started getting powerful enough and cheap enough that it was possible to run Machine Learning algorithms on massive datasets using commodity hardware. One especially important turning point around 2010 was developing effective methods for running Machine Learning algorithms on Graphics Processing Units (GPUs) rather than on the Central Processing Units (CPUs) that handle most computing workloads. Originally designed for video games and computer graphics, GPUs are highly parallelized, which means they can perform large numbers of similar calculations at the same time. It turns out that massive parallelism is extremely useful in speeding up the training of Machine Learning AI models and in running those models operationally. For many types of Machine Learning, using GPUs can speed up the training process by 10-20x while reducing computer hardware costs. Access to the cloud is also very helpful, since organizations can rapidly access massive computing resources on demand (for the relatively short amounts of time needed for training) and limit purchases of computing power to only what they need, when they need it.

3) **Improved Machine Learning Algorithms:** The first Machine Learning algorithms are decades old, and some decades-old algorithms remain incredibly useful. In recent years, however, researchers have discovered many new algorithms that have greatly sharpened the field's cutting-edge. These new algorithms have made Machine Learning models more flexible, more robust, and more capable of solving different types of problems.

4) **Open Source Code Libraries and Frameworks**: The cutting-edge of Machine Learning is not only better than ever, but also more easily available. For a long time, Machine Learning was a specialized niche within computer science. Developing Machine Learning systems required a lot of specific expertise and custom software development that made it out of reach for most organizations. Now, however, there are many open source code libraries and developer tools that allow organizations to use and build upon the work of external communities. As a result, no team or organization has to start from scratch, and many parts that used to require highly specialized expertise have been largely automated. The difficulty of developing an AI model has fallen to the point where – for some applications – even non-experts and beginners can create useful AI tools. In some cases, open source ML models can be entirely reused.

Figure 2. Key Factors Driving Recent Improvements in ML Performance



In short, using Machine Learning generally used to be expensive and complicated, so there were comparatively few applications and organizations for which it was a good fit. Now, however, using Machine Learning is practical and powerful for a far more diverse set of applications. Thanks to the ever-increasing availability of more massive datasets, increased computing power, improved Machine Learning algorithms, and improved open source code libraries and software development frameworks, things that used to be nearly impossible, such as automated facial recognition, are now possible. Programs that used to have terrible performance, such as automatic translation, now have significantly better performance. Finally, AI systems that used to be extremely

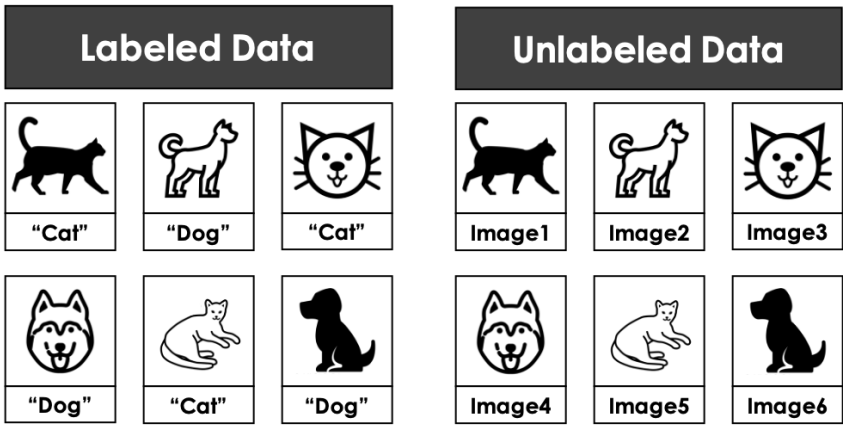
expensive to develop, such as imagery classification, are often now affordable and sometimes even cheap.

Despite their huge potential, AI solutions are not a great fit for all types of problems. If you have an application where you think using AI could be beneficial, knowing whether or not any particular system that is claiming to use "AI" is using Machine Learning is important for several reasons. For one thing, Machine Learning works differently from traditional software, and it has different strengths and weaknesses too. Moreover, Machine Learning tends to break and fail in different ways. A basic understanding of these strengths, weaknesses, and failure modes can help you understand whether or not your particular problems are a good fit for a Machine Learning AI solution.

WHAT ARE THE DIFFERENT TYPES OF MACHINE LEARNING? HOW DO THEY DIFFER?

Like Artificial Intelligence, Machine Learning is also an umbrella term, and there are four different broad families of Machine Learning algorithms. There are also many different subcategories and combinations under these four major families, but a good understanding of these four broad families will be sufficient for the vast majority of DoD employees, including senior leaders in non-technical roles.

Figure 3. Labeled and Unlabeled Training Data



The four categories – explained more on the following page – differ based on what types of data their algorithms can work with. However, the important distinction is not whether the data is audio, images, text, or numbers. Rather, the important distinction is whether or not the training data is labeled or unlabeled and how the system receives its data inputs. Figure 3 provides a simple illustration of labeled and unlabeled training data for a classifier of images of cats and dogs.

Depending upon whether or not data is labeled, a different family of algorithms applies. The four major families of algorithms are Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning.

Supervised Learning: "Supervised" means that, before the algorithm processes the training data, some "supervisor" (which may be a human, group of humans, or a different software system) has accurately labeled each of the data inputs

with its correct associated output. For example, if the goal of the AI system is to correctly classify the objects in different images as either "cat" or "dog," the labeled training data would have image examples paired with the correct classification label. Supervised Learning systems can also be used for identifying the correct labels of continuous numerical outputs. For example, "given this wing shape input, predict the output air drag coefficient."

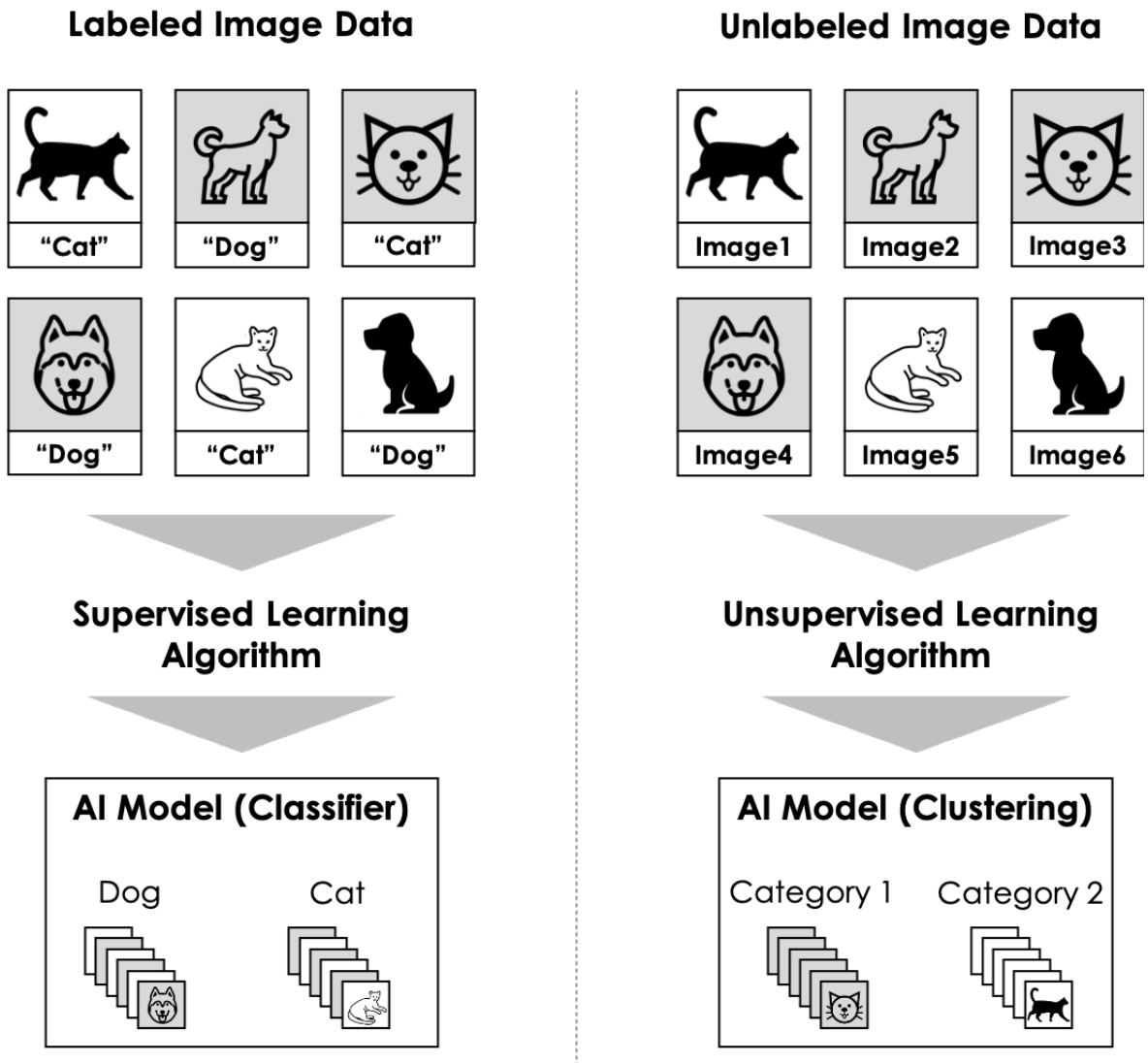
Many Supervised Learning systems can achieve extremely high performance, but they require very large labeled datasets to do so. Using image classification as an example, a common rule of thumb is that the algorithm needs at least 5,000 labeled examples of each category in order to produce an AI model with decent performance. Acquiring all of this labeled data can be easy or very difficult, depending upon the application. In the case of facial recognition algorithms, most companies use paid humans to manually label images. In the case of online shopping recommendation engines, the customers are actually providing the data labels through the normal course of their shopping. The data inputs are the recommended items displayed to the customers and the customer's profile information, while the outputs are the actual purchases made or not made. This is one of the major reasons why internet companies were at the forefront of the adoption of Machine Learning AI: their users were constantly producing valuable datasets – both labeled and unlabeled – and the online environment allowed for rapid experimentation with Machine Learning-enabled analysis and automation.

Note that pre-labeled data is only required for the training data that the algorithm uses to train the AI model. The AI model in operational use with new data will be generating its own labels, the accuracy of which will depend on the AI's training. If the training data set was sufficiently large, high quality, and representative of the diversity present in the operational environment, then the performance of the AI model in generating these labels can be at or above human performance.

Unsupervised Learning: Unsupervised algorithms are those that can extract features from the data without the need for a ground-truth label for the results. Using the aforementioned example of an image classifier, the AI model produced by an unsupervised algorithm would not return that a specific input image was of a "cat" or a "dog." Rather, the model would sort the training dataset into various groups based on their similarity. One sorted group might be the desired groups of cats and dogs, but images might instead be sorted based on undesired categories such as whether or not they have a blue sky in the background, or a wooden floor. Unsupervised Learning systems are therefore often less predictable, but because unlabeled data is almost always more available than labeled data, they remain critical. Additionally, Unsupervised algorithms are very useful when developers seek to explore and understand their own datasets and what properties might be useful in either developing automation or changing operational practices and policies.

Figure 4 depicts the differences between Supervised and Unsupervised algorithms using an image analysis example.

Figure 4. Illustrated Example of Supervised and Unsupervised Algorithms



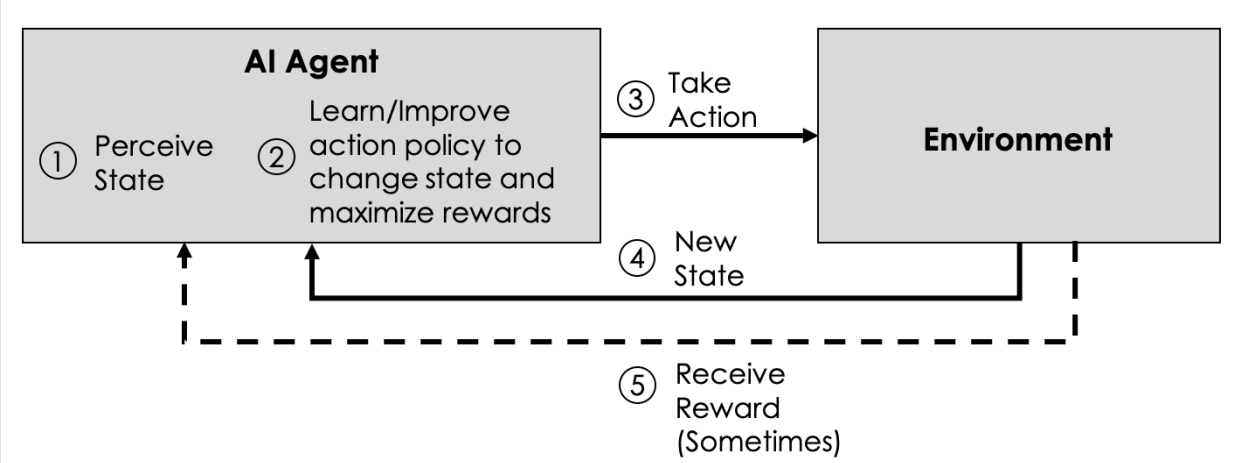
It is not true that Unsupervised Learning is "worse" than Supervised Learning (though performance can be lower for some use cases). Rather, Unsupervised Learning is useful for solving different types of problems. A common Unsupervised use case is fraud detection in financial data. In the case of fraud detection, Supervised Learning could be a good fit for identifying potential fraud that matches behaviors known to be unlawful or associated with fraud. Unsupervised Learning can find new, unidentified patterns of behavior that might indicate new types of fraud techniques.

Semi-Supervised Learning: There is also an increasingly popular class of "Semi-Supervised" algorithms that combine techniques from Supervised and Unsupervised algorithms for applications with a small set of labeled data and a large set of unlabeled data. In practice, using them leads to exactly what you would expect, a mix of some of both of the strengths and weaknesses of Supervised and Unsupervised approaches.

Reinforcement Learning: In Reinforcement Learning, the training data is collected by an autonomous, self-directed AI agent in the course of perceiving its environment (which might be the real world or a simulated environment) and performing goal-directed actions (trying to maximize receipt of "rewards"). Four aspects of Reinforcement Learning are notably distinct from Supervised and Unsupervised Learning:

1) Data is gathered by the AI agent itself in the course of its interacting with the environment and perceiving stated changes. For example, an AI agent playing a digital game of chess makes moves and perceives changes in the board based on its moves.
2) The rewards are input data received by the agent when certain criteria are satisfied. For example, a Reinforcement Learning AI agent in chess will make many moves before each win or loss. These criteria are typically unknown to the agent at the outset of training.
3) Rewards often contain only partial information. A reward like a win in chess conveys that some inputs must have been good, but it doesn't clearly signal which inputs were good and which were not.
4) The system is learning an action policy for taking actions to maximize its receipt of cumulative rewards.

Figure 5: Simplified Reinforcement Learning Diagram



Reinforcement Learning has proven to be very useful in developing high performance AI systems that play games, such as chess, Go, poker, and *StarCraft II*, including systems that can defeat human world-champion-level players. In developing learning systems for these games, there are no labeled datasets that outline every possible move that can be made and provide a true assessment of whether it was a "good move" or a "bad move." Instead, the partial labels only reveal that the final outcome of all the moves made in the game was a "win" or a "loss." Reinforcement algorithms explore the space of possible actions in an effort to learn the optimal policy (set of rules for determining the best action) that will maximize long term rewards.

Reinforcement Learning works very well for games and simulations because the system automatically generates its own training data, which only costs the price of running computational hardware for the algorithm and simulation. For example, AlphaGo, a Reinforcement Learning system focused on the board game Go, played more than 4.9 million games in three days (one full game every nineteen seconds) against itself in order to learn how to play the game at a world-champion level. In real life, a Go game takes ~1 hour.

Reinforcement Learning is currently more challenging to utilize in applications that operate in the real world for three reasons. First, the real world is not as heavily bounded as video games in terms of inputs, outputs, and interactions. Second, time cannot be sped up in the real world. Third, there are consequences to failure in the real world. For example, nothing bad happens when an autonomous driving system crashes in simulation. In the real world, the consequences can be severe, even lethal. As a practical matter, Reinforcement Learning systems have shown significant promise in research settings, but they are much less common than Supervised and Unsupervised Learning applications in real world operations. Nevertheless, many tech companies do use Reinforcement Learning in operational applications, and many researchers expect that real-world usage will grow significantly over the next decade.

Different Types of Machine Learning Summarized

Knowing the different families of Machine Learning algorithms and their strengths and weaknesses is fundamental to making wise decisions about developing and using AI systems. Supervised Learning systems can deliver incredible performance, but acquiring labeled data may be a challenge. Unsupervised Learning systems don't require labeled data, but their performance for some applications will generally be far more limited than Supervised systems. Reinforcement Learning systems can generate their own data but can generally only be used for applications that offer access to simulators that closely resemble the operational environment. Thus far, this includes fewer applications.

When pursuing a specific AI application, developers and program managers should always start by asking the following questions:

- How much training data do we have that is relevant to our desired application? How can we get more?
- How accurate and consistent is the data? How can we increase this?
- Is this data labeled? If not, how difficult would it be to label this data?
- Can only highly trained human experts label the data, or can ordinary people label it?
- Is our training data truly representative of and appropriate for the operational environment? How will this change over time?
- After operational deployment, what types of data changes could cause this system to have degraded performance?
- Do we have access to a high-fidelity simulator that closely resembles our operational environment?

Particular software or hardware products are usually integrated systems with many different elements and functions. One part of a system might use Supervised Learning, while another might use Unsupervised, and so on. In each case, however, those questions will guide the development and management of each subsystem or element of the integrated system.

WHAT ARE NEURAL NETWORKS AND MACHINE LEARNING?

In the mass media, Neural Networks and Deep Learning are two of the most common phrases associated with excitement about modern AI. There's a good reason for that: the dramatic improvements in the performance of AI systems, especially those of recent research breakthroughs, have for the most part been enabled by algorithms that make use of Neural Networks.

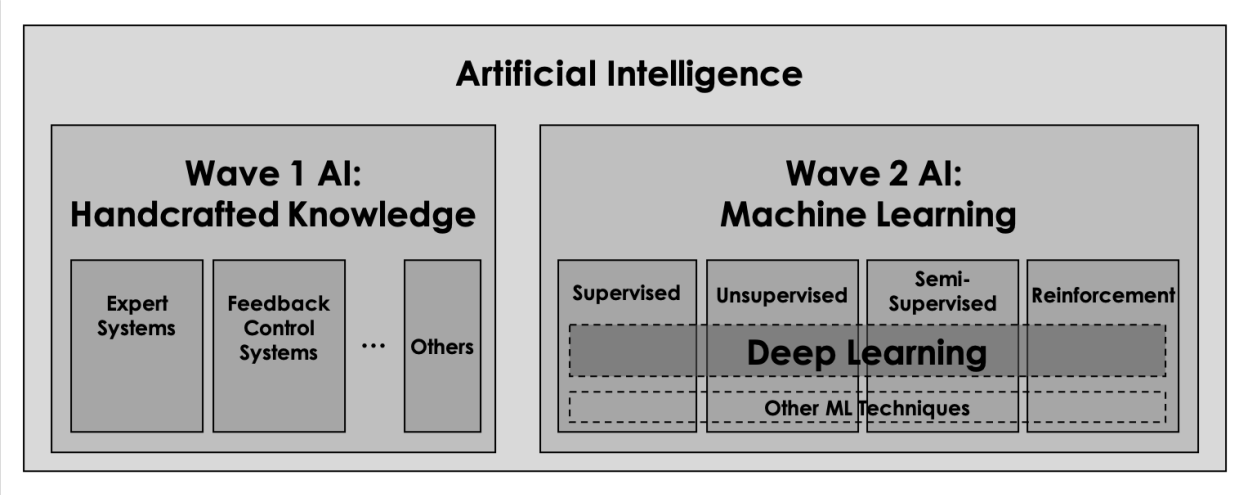
Neural Networks are a specific category of algorithms that are very loosely inspired by biological neurons in the brain. Deep Neural Networks (a.k.a. Deep Learning) merely refers to those Neural Networks that have many layers of connected neurons in sequence ("deep" referring to the number of layers). Though Neural Networks are most strongly associated with Supervised Learning, Deep Learning can, with the right architecture, also be applied to Unsupervised, Semi-Supervised, and Reinforcement Learning.

Neural Networks have been around since the late 1950s, but training Deep Neural Networks only became practical around the 2006 timeframe. Since then, the previously mentioned trends – more data, more computing power, improved algorithms, and improved open source code libraries – have had an especially large impact on improving Deep Learning performance. Since 2012, many of the winning systems in AI competitions around important performance benchmarks

are routinely won by systems that make use of Neural Networks and Deep Learning.

While it is very important for engineers and developers of DoD AI systems and DoD technical leaders to have a good understanding of Neural Networks and how they work, a granular understanding of Neural Networks is overkill and a distraction for most DoD senior leaders. Everything stated earlier in this paper about the general categories of Machine Learning and different types of Machine Learning applies to Neural Networks as well. And knowing whether or not your Machine Learning system is using Neural Networks or another algorithm like decision trees won't have many important implications for how you run your program. In general, AI program managers should care about the performance of the AI system and the types of data required in order to guarantee that performance. For many applications, achieving the required performance will require using Neural Networks. For others, it won't. Focusing on using the most advanced algorithm is important for many parts of the basic research community. For those in positions involving policy, applied R&D, and operations, factors like feasibility, performance, and reliability are more important.

Figure 6. Deep Learning's Place in AI – Using the DARPA "AI Waves" Framework



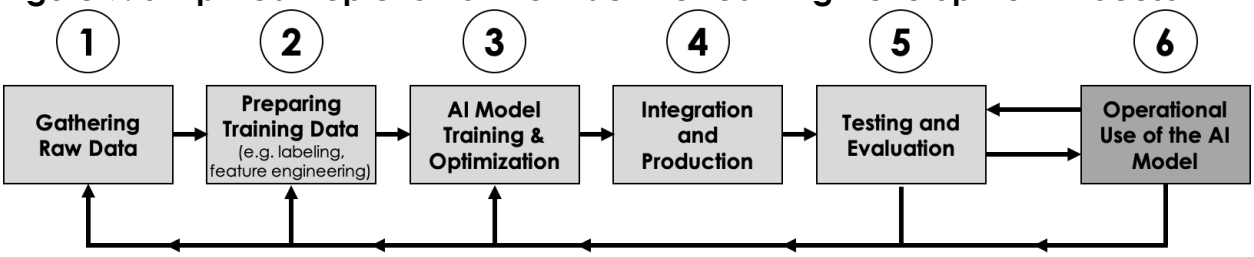
There is one important exception, however. Neural Networks differ from other types of Machine Learning algorithms in that they tend to have low explainability. The system can generate a prediction or other output, and testing can provide evidence suggesting that these predictions have high accuracy, but it is very difficult to understand or explain the specific causal mechanisms by which the Neural Network arrived at its prediction, even for top AI experts. This "explainability problem" is often described as a problem for all of AI, but it is primarily a problem for Neural Networks and Deep Learning. Many other types of Machine Learning algorithms – for example decision trees – have very high explainability.

WHAT ARE THE STEPS OF BUILDING AND OPERATING MACHINE LEARNING SYSTEMS?

By now you should have a good understanding of what the different families of Machine Learning algorithms are and how they work. Figure 7 shows the major steps of actually generating Machine Learning models for operational use.

The ultimate goal of an AI development effort is an AI Model that delivers good performance for a given application at acceptable cost. An AI model built with Machine Learning is generated by providing prepared training data to a suitable AI algorithm. The vast majority of tasks during the AI model training process are performed automatically by the algorithm, but in almost all cases this process will need to be overseen and calibrated by human Machine Learning experts.

Figure 7: Simplified Depiction of the Machine Learning Development Process



The AI model will also likely have to be integrated with existing systems and pass through a suitable testing and evaluation process. Having a high performing AI model by itself, however, is not enough to deliver a positive impact on organizational productivity. The most significant organizational productivity enhancements require not just enhanced technical performance, but also operational processes and staff workflow changes that effectively take advantage of the enhanced performance.

WHAT ARE THE LIMITATIONS OF MACHINE LEARNING SYSTEMS?

AI systems are subject to failures resulting both from accidents (safety failures) and from adversarial malicious activity (security failures). There are many different types of Machine Learning failure modes, but perhaps the most common is when the training data is not sufficiently representative and instructive for the diverse, real-world examples the Machine Learning system will encounter. For example, a satellite imagery classifier that is trained to recognize vehicles exclusively using training data images in a desert environment should be assumed to have degraded performance if the operational data images are of the same vehicles in a grassland, urban, or arctic tundra environment. For the same reason, the performance of ML models in real world applications generally degrades over time if not regularly updated with new training data that reflects the changing state of the world. The software engineering maxim "software is never done" is doubly true for Machine Learning software.

More broadly, there are not yet widely agreed upon safety and reliability standards for the development, testing, and operation of Machine Learning systems. Some methods that have proved critical in ensuring safety and reliability of traditional software – such as formal verification – are not currently available for use on Machine Learning systems. Moreover, some Machine Learning failure modes are not fully understood even at the basic research level. Despite the current challenges, Machine Learning AI systems are already (for some cases) safer and better performing than what they replace. With additional future R&D and improved program management standards, Machine Learning will also be reliably used in a much more diverse set of applications, including safety-critical ones. When this positive scenario is realized, it will not be because AI systems are inherently safe and secure – no technology is – but because the responsible stakeholders took the necessary steps to make AI safe and secure. For the DoD, there are very promising developments on this issue. In February 2020, the DoD officially adopted five Ethical Principles for Artificial Intelligence. The DoD also established an Executive Steering Group with representation from each of the Armed Services and major DoD components to make recommendations for improvements in all aspects of DoD AI Policy and operational usage.

CONCLUSION

Talented human capital and access to AI experts are critical factors for success in DoD's AI strategy. However, the basics of AI technology can be understood by anyone who devotes the time to learn. The concepts in this document provide a technical overview that will be adequate for the vast majority of senior leaders to understand what would be required to adopt and utilize AI for their organization.

Those who want to go further and learn more are encouraged to do so. This document can serve as a useful jumping off point to more advanced and domain-specific subjects. Some recommendations for further reading are provided on the next page. Before moving on, however, readers would be wise to double down and ensure they have a rock-solid understanding of the basics.

REFERENCES AND RECOMMENDED FURTHER READING

For non-technical individuals seeking to better understand Machine Learning

- Launchbury, J. (2017, March 19). A DARPA Perspective on Artificial Intelligence. Retrieved from https://machinelearning.technicacuriosa.com/2017/03/19/a-darpa-perspective-on-artificial-intelligence/
- Karpathy, A. (2017, November 11). Software 2.0. Retrieved from https://medium.com/@karpathy/software-2-0-a64152b37c35
- Domingos, P. (2016). *The master algorithm: how the quest for the ultimate learning machine will remake our world*. New York: Basic books

For engineers seeking to better understand Machine Learning

- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444. doi: 10.1038/nature14539
- Domingos, P. (2012). A few useful things to know about Machine Learning. Communications of the ACM, 55(10), 78–87. doi: 10.1145/2347736.2347755
- Brunton, S., Noack, B., & Koumoutsakos, P. (2020, January 4). Machine Learning for Fluid Mechanics. Retrieved from https://arxiv.org/abs/1905.11075
- Russell, S. J., & Norvig, P. (2010). Artificial Intelligence: a modern approach (3rd ed.). Upper Saddle River: Prentice-Hall. [note: 4th edition expected in mid-2020]
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: The MIT Press.

Helpful context for reading popular media about AI

- Karpathy, A. (2017, May 31). AlphaGo, in context. Retrieved from https://medium.com/@karpathy/alphago-in-context-c47718cb95a5

Organizational best practices for using AI to improve performance

- Berinato, S. (2017, July 19). Inside Facebook's AI Workshop. Retrieved from https://hbr.org/2017/07/inside-facebooks-ai-workshop
- Horneman, A., Mellinger, A., & Ozkaya, I. (2019). AI Engineering: 11 Foundational Practices. Carnegie Mellon University, Software Engineering Institute. https://resources.sei.cmu.edu/asset_files/WhitePaper/2019_019_001_634648.pdf
- Basilico, J. (2017, December 13). Making Netflix Machine Learning Algorithms Reliable. Retrieved from https://www.slideshare.net/justinbasilico/making-netflix-machine-learning-algorithms-reliable
- Kim, G., Debois, P., Willis, J., Humble, J., & Allspaw, J. (2016). *The DevOps handbook* Portland, OR: IT Revolution Press, LLC.

Further Reading on AI Safety and Security

- Department of Defense (2020). DOD Adopts Ethical Principles for Artificial Intelligence. https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/
- Ortega, P. A., & Maini, V. (2018, September 27). DeepMind Safety Research: Building safe Artificial Intelligence https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., & Dan. (2016, July 25). Concrete Problems in AI Safety. Retrieved from https://arxiv.org/abs/1606.06565
- Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are easily fooled: High confidence predictions for unrecognizable images. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2015.7298640